An Entropy and Noisy-Channel Model for Rule Induction

Published by
LOT
Kloveniersburgwal 48
1012 CX Amsterdam
The Netherlands

phone: +31 20 525 2461

e-mail: lot@uva.nl
http://www.lotschool.nl

Cover illustration: Panayotis Mavrakis & Silvia Radulescu.

ISBN: 978-94-6093-392-9
DOI: https://dx.medra.org/10.48273/LOT0607
NUR: 616

An Entropy and Noisy-Channel Model for Rule Induction

Een Entropie en Ruiskanaal Model voor het Leren van Regels
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

vrijdag 12 november 2021 des middags te 4.15 uur

door

Silvia Radulescu

geboren te Ploiesti, Roemenië

Promotoren:

Prof. dr. S. Avrutin
Prof. dr. F.N.K. Wijnen

This research was supported by the Netherlands Organization for Scientific Research (NWO).

In loving memory of my grandmother Elena.

Rule induction happens to us when exposed to language, just like photosynthesis happens to a flower in the sunlight. And this is not a metaphor, but a physical and real process.

Table of Contents

Acknowledgements

I would like to talk about a few people who have been my inspiration and my saviors along the pathway of writing this thesis. I can only extend a brief thank-you here, but my gratitude and love for them escapes these limits that I am confined to.

First and foremost, I would like to thank my physician, Dr. Gijs Limonard, without whose medical professionalism not even a crumb of this work would have come into existence. Gijs treated me with utmost care and soothing kindness for one year, and he saved my life from deadly bacteria looming inside my lungs. Thank you, Gijs, once more, for your dedication and your kindness. Whenever I am lost, I remember your words "Silvia, don't worry, you will be just fine!"

Rocío Romero Mérida, my sweet princess, without you taking care of me during those times of illness, when I was all alone in a foreign country, without family or friends, I would not have been able to make it one day to write this thesis. Thank you for taking care of me and loving me.

And now to my mentors...

Many years later, as she was sitting down to write these very words, Silvia was to remember that sunny afternoon in Utrecht when she met Sergey Avrutin. At that time, she was a mere Research Master student, just like she'll always be, thirsty for knowledge, and Sergey was teaching a course in "Neurocognition of Language". That world was so recent for Silvia, that many concepts lacked names in her mind, or many names lacked concepts…cannot make up my mind which of the two. Sergey was like a painter of a brand new world, and I was listening in fascination. After those four hours of class, I went walking for four kilometers to my dorm, and couldn't stop thinking of all those mesmerizing ideas and studies that Sergey talked about. I don't remember ever having felt such unbridled enchantment. That enchantment was the cradle of all my future research endeavors, and the motivation for this dissertation. Thank you, Sergey, for being my inspiration ever since. Thank you for guiding my thinking with so much patience and enthusiasm. Thank you for listening to all my crazy thoughts and ideas about entropy and language, about thermodynamics and memory and cognition, and for being so excited about everything I was raving about. The last ten years of my life have been much better thanks to you.

A few months later, as she was sitting in the classroom sipping her coffee to keep herself from not falling asleep (the illness was already taking the best of her), Silvia was to remember how it felt to face the devil's advocate. Frank Wijnen was the professor teaching the "Language. Speech. Brain" course of the Research Master program, and he asked her a question that almost made her choke on her coffee. Frank was determined to doubt and question every single word Silvia was going to utter. Challenge accepted! Thank you, Frank, for always listening to all my ideas and always questioning them to the bone, until I managed to polish them into something readable. I still don't know if I have ever convinced you of anything that is written in this thesis. I don't need to know, the thrill of trying to was all worth it. Now, I have developed this weird habit of questioning all my thoughts.

I would like to thank the members of the assessment committee of professors who read and approved this dissertation: Jos van Berkum, René Kager, Padraic Monaghan, Daniel Swingley and Ramon Guevara Erra. Each one of you has been there at some decisive moment on my way to wisdom and knowledge.

Jos, I thank you deeply for the marvelous discussions and for all the amazing books you have recommended to me (I am still reading through Antonio Damasio's books, from time to time). Your teachings about the emotion system helped me cope with everyday life. Thank you also for giving me valuable feedback on the research proposal I submitted to the Netherlands Organization for Scientific Research, which awarded me the grant for this dissertation.

René, besides teaching me phonology in my Master program, I am also greatly thankful to you for helping me with valuable feedback not only on the essential stuff (my research proposal for this project), but also on the small stuff (like listening to and evaluating the intelligibility of the Dutch stimuli for the experiments of Chapter 5).

Very special thanks go to Padraic Monaghan, who showed interest in my work and encouraged me from my first talk in an international conference, in Valetta in 2015. I was a total newbie back then, much more than now, still you chose a very nice way to ask me a difficult question after my talk and after my obviously rather poor answer, you still gave me a great encouraging smile. All our brief conversations and your scientific work have been a great inspiration to me. I have tried my best to learn from you, and…I will not hide it, I have tried my best to take after your fabulous writing skills.

When my young researcher wings had been clipped and my idealistic picture of scientific publication was shattered, Daniel Swingley was there to restore my trust and self-esteem by helping me publish my first paper (Chapter 1 of this dissertation). Thank you, Daniel, your scrutinizing, yet inspiring questions on my manuscript helped me enormously in shaping and expressing my ideas.

Ramon Guevara Erra, when I met you and had our first conversations about entropy and thermodynamics, and after I read a couple of your papers, I finally believed it: I wasn't delusional to have those ideas about entropy and the brain. At least, I found out that I wasn't the only one who thought the conscious (awake) brain craves for entropy, not for less entropy, as the current dogma holds. When you said you actually found evidence for that in your research, I knew I had to write my thoughts down. And here they are in Chapter 7. Thank you for that and for the many discussions and all the literature that you introduced me to.

I will never forget all the constructive debates I had during the Syntax classes with Riny Huijbregts. Riny, your class hand-outs were the best in the history of the Research Master program, and your passion for Chomskyan syntax was contagious even for a heretic like me. Not only were you an incessant inspiration for me, but one of the best pedagogues I have ever encountered (and we do have especially great educators in Romania). Thank you, really, thank you.

I also want to thank Maaike Schoorlemmer for believing in me and supporting me throughout the last ten years. Maaike, thank you from the bottom of my heart for being there for me and encouraging me when I was ill during the Master's program, and also for all your patience in answering the tens or maybe hundreds of questions I asked you. Also, thank you for not scolding me for never being in the office.

And now to my colleagues and other professors who helped me…

First and foremost, I need to thank my colleague and mentor Ileana Grama. Ileana, thank you for entrusting me with running one of the experiments in your dissertation, as part of my internship, although it was my very first experiment and I could have messed things up real good. Thank you for all our fruitful conversations about science and life, either in Utrecht or while having pinchos in San Sebastian or Bilbao. I am very happy with our collaboration, which has resulted in two chapters of my dissertation (Chapter 4 & 5). Looking forward to our next endeavor!

I would like to thank all my colleagues from the lab at the Utrecht University, starting with Brigitta Keij. Brigitta, thank you for all your fruitful comments on the early results of my experiments, and also for being so kind and patient to help me record all the Dutch stimuli that I used in all the XXY experiments in my dissertation.

I am very grateful to Iris Mulders, who has patiently set up all my experiments, and they were many…so many that at some point she was having trouble finding new participants who hadn't already participated in any of my experiments. I am also very thankful to Kirsten Schutter for teaching me a great deal about statistics and for helping me analyze the data of all my experiments. Thank you, Kirsten for all the statistics workshops you organized at the University, for all our long conversations about statistics. You made statistics much more digestible.

I cannot speak about statistics without thanking my professor of statistics Hugo Quené. Hugo, I was there in all your classes listening carefully to all your teachings. Afterwards, whenever I was reading a scientific study or running my own, I would always remember your first lesson: you showed us a very cute picture of what looked like a mommy monkey sweetly holding her baby monkey. You asked us to write down very concisely what we saw in the picture. Many wrote "An artistic depiction of animal motherly love." Or "Sweet mommy monkey holds her baby before going to sleep.". I wrote drily and cold-heartedly: "In the foreground of the picture, a brown monkey holds a small monkey in a field of grass with many shades of green. The background is blurred, due to the depth of field technique used." You approved of my story, and you taught us to not make assumptions and to not look for interpretations in our data which might not be there, to stay focused only on what is clearly there in the data.

I would also like to express my gratitude to Henriette de Swart for giving me the amazing opportunity to co-teach the Research Seminar in the Research Master program for two years in a row. I learned so much from this experience, and I want to thank you deeply for all our conversations about teaching and about academia, in general. I am also very thankful to Aoju Chen for her interest in my work and our conversations regarding my fNIRS study.

I am very thankful to a lot of colleagues at Utrecht University for our conversations and for their valuable comments on my work: Carolien van den Hazelkamp, Marjolein van Egmond (indeed, we should have had more conversations about science), Maartje de Klerk, Andrea Santana Covarrubias, and all the people who participated in the lab meetings on all those Friday mornings, when everybody looked so much more awake than me…

I am very thankful to all my office mates from Utrecht University who had to put up with the desolating image of my empty desk throughout all my PhD years: Shuangshuang Hu, Jorik van Engeland, Heidi Klockmann, Emma Everaert (special thanks for the translation of the Summary of this dissertation in Dutch).

Special thanks go to my dear Shuangshuang Hu, whom I have met more often in the lab in Janskerkhof 13, where I was running my experiments or she was running hers. Thank you, Shuangshuang, for all the hugsssss, and for being so smiley and shiny all the time. I enjoyed all the times we spent together whether we were organizing huge and essential events like Uiltjesdagen 2016, or we were going to my salsa parties, or making plans to play badminton, to meet in Paris or Tokyo (none of which came true …). You were my bright sunshine in the gloomy basement lab of Janskerkhof, you made my time there less tedious.

I am also very thankful to my two research assistants in the babylab at the University Descartes in Paris, Annie Gaule and Anthony Picaud. Thank you for making my work there much more enjoyable, and for helping me tame those French babies and parents into going through my brain imaging experiments (Chapter 2).

And now to all my students and interns…

I really owe a huge thank you to all my students and interns who helped me learn and accomplish so much more than I would have done on my own. Efi Giannopoulou, thank you for running one of my experiments (in Chapter 3), and for all your hard work on posters and talks that we presented together in international conferences. I felt honored when you got inspired by my ideas and you decided to run a study (for your master thesis) from a different angle on the entropy model, namely the effect of mood on rule induction. Jessica van Schagen, thank you for running a pilot experiment that explored the pristine field of the cognitive capacities underlying channel capacity (your study was the pilot that facilitated the design of the experiments presented in Chapter 6). Mridhula Murali, thank you very much for all your patience and hard work on one of my experiments from Chapter 6. I enjoyed our collaboration very much and I haven't forgotten about visiting you in that amazing city where you're currently doing your PhD – Edinburgh. James Pickett, thank you very much for the courage to embark on an internship/thesis based on a vague topic: the predictions of my entropy model at the acoustic level. Although, the results were inconclusive, I enjoyed every bit of our research and all our long conversations related to entropy and rule induction.

Very special thanks go to Areti Kotsolakou, who ran both experiments presented in Chapter 5 as part of her internship and analyzed the data. Thank you also for your valuable feedback on my Chapter 5 and for collaborating on submitting a manuscript for publication based on that chapter. Your hard work and dedication, your brilliant critical thinking and your enthusiasm are rare. I feel very happy and honored that you continued your career with a PhD project inspired by the entropy model, and I am sure your research will add great value to the entropy-related perspective on rule learning.

I would also like to thank all my Research Master students from the Research Seminar that I had the honor to teach in 2017 and 2018. The list would be way too long to include here, but I would like each one of you to know that I have learned so much with you and from you while teaching you the little bit that I know.

And now to my family and friends…

Many years later, as she was studying physics books for this dissertation, Silvia was to remember that mystical day when she discovered thermodynamics. She was four years old, and she asked her daddy the following question: "You take a bottle of water out of the fridge, put it on the table, it gets warmer. You take a cup of tea, put it on the table, it gets colder. What is happening on the table?" Vasile – a physics lover – was excited like a kid, he was about to explain thermodynamics to his 4-year-old daughter. Daddy, I thank you from the bottom of my heart for answering that question and all the other million questions about physics and bridge construction that followed throughout the years. The thought of dedicating these lines to you kept me going through the hardships of writing this dissertation.

I am also very grateful to my mother, Georgeta, for always pushing me to do better in life, to study more, to whine less and to never give up. Thank you for raising me to be strong, respectful, good-mannered and happy with what I have in life. You inspired me to read all those thousands of books from your collection, so that today I can offer you a book written by myself. Thank you for teaching me math and introducing me to the greatest band of all times – Queen.

Twenty years after the last time I looked into your beautiful green eyes, my beloved granny, I can still feel your sweetness and your cosy touch, your unconditional love, your dedication and patience. You are the reason I am. I cannot thank you enough for everything you have done for me. Grandma Elena, I will always love you and miss you.

Heartfelt thanks also go to my brother, Marian, who has taken a special interest in my research. Thank you for giving me extensive feedback on my first manuscript of Chapter 1, and for being so keen on hearing about the results of my research. Your brilliant questions and input helped me understand much better what I wanted to express, and gave me a brand new insight into my research from the perspective of a software engineer. Although, we don't have a chance to meet often because we live so far apart from each other, I want you to know that I always think of you dearly, and I feel you as a great support in my life, both professionally and personally.

I was very lucky to have my lovely friends by my side, always supporting me. Teodora Dumitru, thanks for all our deep conversations about life, literature, science, and for introducing me to the constructal law. Irina Chiscop, thank you for listening to all my long stories, and for acting normal when your husband stumbled upon us in the kitchen and was totally flabbergasted by our passionate discussion about entropy instead of some girl talk. Thank you, Liviu Radulescu, for your precious help with coding the scripts I used for running most of the experiments in this dissertation and for your brilliant insights on my research.

I could not have made it through all these years of hard work without my salsa friends, with whom I shared countless dances and fun moments, while experiencing our passion for music and dance. Mark Martis, thank you for talking me through some of the hardest times in my life and for all the amazing dances. Maarten Jansen, thank you for helping me and making my stay in the Netherlands more enjoyable, and for all the fantastic salsa music you introduced me to. Pedram Sadriani, thank you for all the breathtaking salsa dances that kept me alive and in the best shape. Florina Vasie, thank you for being by my side and for all the fun we had together.

Finally, I want to thank my Cretan friends and family who took very good care of me throughout all these years: Elli Makridaki, Antonia, Konstandina and Nikos Iliakis. Thank you, Evangelia Zacharioudaki, Ioannis Mavrakis and Elpida Mavraki, I am truly grateful to you for welcoming me into your wonderful and loving family, and for giving me a sweet home in your hearts.

Last but not least, I want to thank Panayotis Mavrakis, my other half, my soulmate. Thank you for being by my side while I was writing this dissertation and putting up with me when I wasn't being myself because of the pressure. Thank you for all the fabulous late-night conversations about my research, and for making my creepy ideas come to life in an artistic and cute way on the cover of this dissertation. I treasure every single moment we spent together, all the laughs and all the tears we shared. You were there for me to tame my insomniac mind and guard my sleep, to bring me up when I was feeling down, to soothe my troubled mind with your magic. I am very fond of how we butchered English, Romanian and Greek into our own special pidgin. You taught me how to enjoy the Cretan ways of life, how to eat all the *skepasti* without feeling any guilt, and you introduced me to music theory and guitar playing. I am honored to be your number one fan girl until the end of time. Thank you for your organic kindness, your tenderness, your unconditional love, your honesty, your respect, your everything. I love you deeply.

General introduction

This dissertation is a collection of articles (published, under review or to be submitted for publication), which present the results of a research project that investigated linguistic rule induction from an information-theoretic perspective. The main goal of this research project was to propose and test an innovative entropy model for rule induction based on Shannon's *noisy-channel coding theory* (Shannon, 1948).

Rule induction (generalization or regularization) is an essential language acquisition mechanism that empowers language learners to not only memorize specific items (e.g. phonemes, words) experienced when exposed to linguistic input (language), but also to acquire relations between these items. These relations range from statistical patterns between specific items present in the linguistic input (Saffran, Aslin, & Newport, 1996; Thiessen & Saffran, 2007) to more abstract category/rule induction (Marcus, Vijayan, Rao, & Vishton, 1999; Smith & Wonnacott, 2010; Wonnacott, 2011; Wonnacott & Newport, 2005). For example, language learners not only memorize words and combinations of words, like *mom walked slowly* and *dad talked nicely,* but they also infer generalizations (rules) like 'add *-ed*' or 'add *-ly*' to specific items in order to express a past action or the manner of carrying out an action. Moreover, learners' rule induction abilities also move away from specific combinations of items to abstract categories and generalized rules: for example, Noun-Verb-Adverb is a well-formed sequence, where each category can take a virtually infinite number of specific items. This research project addressed the inductive steps from memorizing specific items and combinations of items, to inferring rules (or statistical patterns) between these specific items, and also to forming categories and generalized rules that apply to categories of items.

Following definitions from previous literature (Gómez and Gerken, 2000), we distinguish between two forms of rule induction: *item-bound generalizations* and *category-based generalization. Item-bound generalizations* describe generalizations (rules) bound to specific physical items present in the experienced input (e.g. a relation based on physical identity, like "*la* follows *la*" or "add a specific item *-ed"*). Conversely, *category-based generalizations* are operations beyond specific items from the input, spanning over a virtually infinite number of novel instances. They describe relations between categories/variables, for example, "Adverb follows Verb" in a Verb-Adverb pattern, where Verb and Adverb are categories (variables) taking different specific items as values, for example, *"walk", "slowly",* etc.

Previously, two factors were shown to modulate rule induction: either the variability of items experienced by the learner – i.e. input variability (Gerken, 2006; Gómez, 2002; Reeder, Newport, & Aslin, 2013), or certain limited cognitive capacities that support the encoding of the input, e.g. memory capacity (Hudson Kam & Chang, 2009; Hudson Kam & Newport, 2009; Newport, 1990; 2016). However, the underlying mechanism and the exact dynamics between these two factors that drive the inductive steps from memorizing specific items and statistical regularities to inferring abstract rules remain largely underspecified. The articles presented in this dissertation aim at filling this gap.

The mechanisms underlying rule induction have been the object of a heated ongoing debate in psycholinguistics. On the one hand, there is mounting evidence that an item-bound mechanism, which relies on memorization of specific items and statistical computations about their probability distribution, namely *statistical learning,* would suffice for encoding the input by *item-bound generalizations.* For example, phonotactic information (Chambers, Onishi, & Fisher, 2003), and word boundaries (Aslin, Saffran & Newport, 1998; Saffran, Aslin & Newport, 1996) were shown to be acquired by basic statistical computations, such as transitional probabilities (i.e. the probability of a specific item occurring after another). On the other hand, other researchers argued that statistical learning alone cannot account for generalizations beyond specific items evidenced by the input (Endress & Bonatti, 2007; Marcus, Vijayan, Rao, & Vishton, 1999). Thence, they proposed another mechanism for encoding the input as *category-based generalizations,* namely *abstract (algebraic) rule learning.* More recently, a *single-mechanism hypothesis* was put forth, according to which the same mechanism – *statistical learning* – underlies both item-specific and abstract learning (Aslin & Newport, 2012; 2014; Frost & Monaghan, 2016). While supporting the *single-mechanism hypothesis,* this dissertation aims at better specifying the nature of this mechanism and *how* one single mechanism leads to these two qualitatively different types of generalization.

While previous studies used this terminology to refer both to the processes (mechanisms), and the two forms of encoding/generalizations (statistical regularities vs abstract rules), this dissertation takes a step further, and proposes that, the mechanism(s) underlying rule induction should be conceptualized separately from their outcomes, that is from the resulting forms of encoding (*item-bound generalizations* and *category-based generalizations*). This distinction allows for the main research questions of this research project to be formulated:

1. Are the two forms of encoding outcomes of two separate mechanisms, with *statistical learning* resulting into the lower-level *item-bound generalizations*, and *abstract rule learning* outputting the higher-order *category-based generalizations*?

2. Or, are these forms of encoding outcomes of the same mechanism?

2.a. If this is the case, is it a phased mechanism that *gradually* moves from one form of encoding to the other?

2.b. Or is it an *abrupt* shift?

3. What drives the change in form of encoding from *item-bound* to *category-based generalization*, be it a gradual transition or an abrupt shift? In other words, what are the factors that drive rule induction?

In order to answer these research questions, this dissertation puts forth a novel entropy and noisy-channel capacity model (in short, entropy model) for rule induction, which is based on Shannon's *noisy-channel coding theory* (Shannon, 1948). In short, and simplifying for now, *entropy* is an information-theoretic measure of information, while *channel capacity* is the amount of information (including the noise) that can be transmitted per unit of time. Shannon's coding theory says that a message (i.e. information) can only be transmitted reliably (i.e. with the least loss of information), if encoded by using an efficient encoding method. An encoding method is efficient (reliable), if and only if the rate of information transmission (i.e. entropy per second), plus the *noise*, is below the channel's capacity. If the rate of information transmission is higher than the *channel capacity*, then another more efficient encoding method can be found, but the *channel capacity* cannot be exceeded. If the *channel capacity* is exceeded, there will be loss of information, which renders the encoding method inefficient.

The main hypothesis of the entropy model is that rule induction is an encoding mechanism *gradually* driven by the dynamics between an external factor – *input entropy* – and an internal factor – *channel capacity. Input entropy* quantifies (in bits of information) the statistical properties of the linguistic input, given by the number of items and their probability distribution. *Channel capacity* is used as an information-theoretic measure of the encoding capacity of our brain, and is determined by the amount of entropy that can be encoded per unit of time. In other words, we define our brain's encoding capacity as *channel capacity* (at the computational level, in the sense of Marr (1982)), which is the finite rate of information encoding (bits per second). At the algorithmic level, the *channel capacity* might be supported by several cognitive capacities involved in processing and encoding information, e.g. memory capacity and attention.

Among other studies that used entropy measures to look into generalization (or regularization) patterns (Ferdinand, 2015; Ferdinand, Kirby, & Smith, 2019; Perfors, 2012; 2016; Saldana, Smith, Kirby, & Culbertson, 2017; Samara, Smith, Brown, and Wonnacott, 2017), this dissertation takes a step further and proposes an information-theoretic model that captures the dynamics between the *input entropy* and our encoding capacity, i.e. *channel capacity*.

Our proposal that rule induction is driven not only by external factors, like input variability, but also by internal factors, like the relevant cognitive capacities, is closely related to another line of research – the classical *Less-is-More hypothesis* (Newport, 1990; 2016), which looks into rule induction in terms of cognitive constraints on learning. According to this hypothesis, overloading our limited memory capacity leads to difficulties in storing and retrieving low-frequency items, which prompts overuse of more frequent forms leading to generalization. These limitations on cognitive capacities, which develop with age, were proposed to explain why young learners have a higher tendency to generalize than adult learners (Hudson Kam & Newport, 2005; Hudson Kam & Newport, 2009; Newport, 1990; 2016).

The entropy model proposed in this dissertation offers an extended and more refined information-theoretic approach to the *Less-is-More hypothesis*, by bringing together both factors (input entropy and cognitive capacity) in one formula. This model takes a step further from the algorithmic level – (i.e. from the cognitive limitations of the memory and attentional resources) – to the computational level, i.e. *channel capacity* – our time-dependent noisy information processor.

In order to answer the research questions, the entropy and noisy-channel model presented in this dissertation puts forth the following hypotheses:

1. *Item-bound generalization* and *category-based generalization* are not separate mechanisms. Rather, they are outcomes of the same encoding mechanism – computationally – which is statistical in nature, and can be captured under Shannon's *noisy-channel coding theory* (Shannon, 1948), as follows.

2. Rule induction is a phased mechanism that moves *gradually* from memorizing specific items and combinations of items, to a high-specificity form of encoding – *item-bound generalization* – and to a high-generality form of encoding – *category-based generalization.*

3. The gradual transition from one form of encoding to the other is driven by the dynamics between *input entropy* and *channel capacity*:

a. Little *input entropy* – that is below or matches the *channel capacity* – facilitates memorizing and encoding specific items, with their physical features, and relations between them, i.e. *item-bound generalization*.

b. An *input entropy* that is higher than the *channel capacity* drives *category-based generalization*, as a more efficient form of encoding.

Specifically, in information-theoretic terms, if the *input entropy* is below or matches the *channel capacity* (i.e. the maximum amount of input entropy that can be encoded per second), the information about specific items and relations between them can be encoded with a high-fidelity item specificity by *item-bound generalization*, at the channel rate (i.e. *channel capacity*). If the *input entropy* is higher than the *channel capacity*, that is the number of items and their probability distribution reach a complexity that is higher than the encoding capacity, then the high-specificity form of encoding – *item-bound generalization* – becomes prone to errors, due to *noise* interference. The information cannot be encoded reliably, and so the form of encoding becomes inefficient, due to loss of information. Thus, the form of encoding is *gradually – bit by bit* – shaped into a high-generality form of encoding – *category-based generalization*, in order to avoid exceeding the *channel capacity.*

In order to test these hypotheses of the entropy and noisy-channel model, we follow previous research on rule induction and we employ the artificial grammar learning paradigm (Gerken, 2006; Gómez, 2002; Marcus et al., 1999; Reeder, Newport, & Aslin, 2013). Also, we use a repetition-based type of grammar – XXY – similar to previous related studies (Gerken, 2006; Marcus et al., 1999), but also a more complex non-adjacent dependency aXb grammar similar to the one employed by Gómez (2002).

The first study of this dissertation (Chapter 1) lays the foundation for this entire research project by formulating the main research questions and by introducing the entropy and noisy-channel model. Chapter 1 probes the effect of the first factor of the entropy model – *input entropy* – on rule induction with adults exposed to a repetition-based XXY grammar (e.g. *da:-da:-li*). Specifically, rule learning is hypothesized to be a phased encoding mechanism that starts out by memorizing specific items (e.g. phonemes, syllables, words) and the statistical regularities between them as instanced in the input, which lays the basis for inferring generalizations between specific items from the input, i.e. *item-bound generalization*. In the case of an XXY grammar, *item-bound generalization* means inferring a *same-same-different* rule only with familiar syllables from the experienced stimuli. An increase in entropy is hypothesized to drive the tendency towards a more abstract *category-based generalization* (i.e. a *same-same-different* rule with novel syllables, as well). Thus, learning this type of XXY grammar entails moving away from *item-bound generalization* to *category-based generalization*, which is predicted to be *gradually* driven by increasing *input entropy*.

In order to test this hypothesis, in two artificial grammar experiments, adults are exposed to a 3-syllable repetition-based XXY artificial grammar (e.g. *da:-da:-li*), in six experimental conditions with increasing *input entropy*. We propose an innovative method to calculate entropy in an artificial grammar, by applying Shannon's entropy formula (Shannon, 1948) to calculate a bigram/trigram average entropy. This method extends and fine-tunes a similar proposal by Pothos (2010).

This chapter is a slightly modified version of a published article – Radulescu, Wijnen, and Avrutin (2019), which shall be referenced as such henceforth.

In Chapter 2, we extend the entropy model in order to address a developmental research question that is motivated by the interaction between *input entropy* and *channel capacity*, as predicted by the entropy model. Since the *channel capacity* is supported by cognitive capacities that develop with age, like memory, infants are hypothesized to have a reduced *channel capacity* compared to adults. Thus, infants' tendency towards rule induction is predicted to be driven by less *input entropy* than the adults. Specifically, we address the question of *whether* and *how* infants process and encode a repetition grammar (ABB) as compared to a non-repetition grammar (ABC), and weather *input entropy* has an effect on this process.

To this end, we test whether and how six-month-old infants process repetition-based linguistic regularities (ABB, e.g. "bu ra ra") as compared to non-repetition sequences (ABC, e.g. "bu fa zo"), by using functional near-infrared spectroscopy (fNIRS), and we manipulate the *input entropy* (low vs high). Infants are predicted to be able to process both ABB and ABC sequences, and also to discriminate between these sequences, while doing so more readily under conditions of high entropy.

In Chapter 3, the conceptual distinction between *item-bound generalization* and *category-based generalization* is defined in more detail, and the transition from one to the other, as an effect of *input entropy,* is probed in another experiment with adults. The main goal of the study is to further investigate the *gradual* transition from rote memorization to *item-bound generalization* and *category-based generalization*, as hypothesized by the entropy model.

Specifically, we expose adults to a low and a medium entropy version of the XXY grammar (from Chapter 1), and we test the hypothesis that low input entropy facilitates not only rote memorization of specific items and their probability distribution evidenced by the input, but also *item-bound generalization*. We also look at individual differences in specific components of cognitive capacities that we hypothesize to underlie *channel capacity*, i.e. explicit/implicit memory capacity and a general-domain pattern recognition capacity, which draws on working memory resources.

Chapter 3 ends with a detailed discussion of our findings and our entropy model in terms of their contribution to the ongoing debate between the two prominent views on the mechanisms underlying rule induction: the *more-mechanisms hypothesis* (Endress & Bonatti, 2007; Marcus et al., 1999) vs the *single-mechanism hypothesis* (Aslin & Newport, 2012; 2014; Frost & Monaghan, 2016).

In Chapter 4, we further extend our entropy model for rule induction from the repetition-based XXY grammar to a more complex non-adjacent dependency a_iXb_i grammar. In this type of grammar specific items *a* always predict specific items *b,* and create frozen $a_i\_b_i$ frames over a richer intervening category of Xs. We argue that learning such a complex type of grammar entails both *item-bound generalization* (a generalization of the dependency between specific *a* and *b* elements – the specific $a_i\_b_i$ frames), and *category-based generalization* (generalizing the intervening category of Xs). This type of grammar poses a challenge to the entropy model, in that successful learners of this type of a_iXb_i grammar move away from an *item-bound* to a *category-based generalization* for the intervening X category, while, crucially, sticking to an *item-bound generalization* for the specific $a_i\_b_i$ dependencies.

We hypothesize that, while high *input entropy* drives *category-based generalization* for the X category, it impedes *item-bound generalization* for the specific $a_i\_b_i$ dependencies of an a_iXb_i grammar. Hence, the effect of increasing entropy on learning this type of grammar is not a *gradually* better performance as we found for the XXY grammar (Radulescu et al., 2019). Rather a U-shape learning curve is predicted, with either low or high *input entropy* (i.e. a lower and an upper bound determined by the *channel capacity*) being expected to facilitate detection of the specific $a_i\_b_i$ dependencies and generalizing them over the category of intervening Xs.

Our entropy model takes a step further from previous theories that claimed the set size of the intervening Xs plays a crucial effect on non-adjacent dependency learning (Gómez, 2002; Gómez & Maye, 2005), namely, a large set size of the middle X elements was proposed to support better learning. We hypothesize that is not mere set size, rather it is *input entropy* that modulates learning. Thus, we aim at teasing apart the effect of set size from the effect of *input entropy*, by keeping a large set size of intervening Xs constant and varying the probability distribution of the items to obtain three different *input entropy* conditions.

To this end, we expose adults to three entropy conditions – low, medium and high – of a non-adjacent dependency a_iXb_i grammar similar to the one by Gómez (2002). If indeed the factor at stake is input entropy, then high *input entropy* is expected to drive better learning than low and medium entropy, in spite of a constantly large set size. Moreover, low entropy is hypothesized to facilitate remembering specific items and relations between them, leading to detection of the non-adjacent $a_i\_b_i$ dependencies, while high entropy is expected to drive generalization of the middle X elements, which also supports non-adjacent dependency grammar learning. Conversely, medium entropy is predicted to create an uncertain environment, which tampers with *item-bound generalization* and does not drive *category-based generalization* either, such that non-adjacent dependency grammar learning is impaired.

This chapter is a slightly modified version of an article in review – Radulescu and Grama (2021), which shall be referenced as such henceforth. In Chapter 5, we probe the effect of the internal factor of the entropy model – *channel capacity* – on rule induction with adults. This factor adds into the "formula" for rule induction the crucial dimension of *time*. According to Shannon (1948), a message (or information) can be transmitted reliably (i.e. with the least loss of information), only if it is encoded by a method that is efficient enough, so that the rate of information transmission (bits per unit of time), including noise, is below the channel's capacity. In short, *channel capacity* is defined mathematically as the maximum possible rate of information transmission, which can be achieved only if the encoding method is adequate and efficient.

Based on these concepts, our entropy model for rule induction hypothesizes that what drives the change in the encoding method – from *item-bound* to *category-based generalization* – is a regulatory mechanism that moves from an inefficient encoding method to a more efficient one, in order to avoid exceeding the *channel capacity*. Efficiency of encoding method means the least loss of information caused by the noise interference during transmission through the channel in time. Thus, reduction of time increases the rate of information transmission and brings a higher inflow of noise, which interferes with information transmission and causes an increased loss of information. This drives the change from *item-bound* to *category-based generalization*. Hence, we hypothesize that it is precisely the finite *channel capacity* that drives restructuring of the information to find a more efficient encoding method.

In order to probe this hypothesis, in Chapter 5 we first show, theoretically, how *channel capacity* and the rate of information transmission can be estimated in an artificial language learning environment for rule induction, namely using the results of our experiments from Radulescu et al. (2019). Next, we directly manipulate the time variable of the *channel capacity* in two other artificial grammar experiments with adults.

To this end, we speed up the bit rate of information transmission, crucially not by simply reducing the time between stimuli by an arbitrary amount, but by a factor that we calculated based on data from our previous experiments (Radulescu et al., 2019), by using the *channel capacity* formula. In the first experiment, we expose adults to the lowest entropy version of the XXY grammar from Radulescu et al. (2019), either in a slow rate of transmission condition or a fast rate of transmission condition. In the second experiment, we expose adults to a low entropy condition of the a_iXb_i grammar (from Chapter 4), one group to a slow rate of transmission and another group to a fast rate of transmission.

We also control for individual differences in specific components of cognitive capacities that we hypothesize to support *channel capacity* at the algorithmic level, i.e. explicit/implicit memory capacity and a general-domain pattern recognition capacity, which draws on working memory resources.

This chapter is a longer version of an article in review – Radulescu, Kotsolakou, Wijnen, Avrutin and Grama (2021), which shall be referenced as such henceforth.

In Chapter 6, we further test the model by looking into the effect of the noisy-channel capacity (Shannon, 1948), by adding *noise* (i.e. random stimulus-irrelevant material) in the background of the lowest entropy version of the XXY grammar task from Radulescu et al. (2019). According to Shannon's noisy-channel coding theory, in a communication system, information can be transmitted reliably, if and only if encoded by an encoding method which is efficient in such a way that the rate of information transmission, including *noise*, is below the channel's capacity.

The efficiency of the encoding method is defined by the ratio of the rate of transmission to the capacity of the channel. Thus, we hypothesize that noise adds sufficient entropy per second, in order to drive a change towards a more efficient encoding method. Specifically, *noise* inflow perturbs the signal and increases the loss of information. Since *noise* increases the loss of information, which in turn calls for a more efficient encoding method, we expect that an increase in *noise* should accelerate the drive towards a reorganization of the information, such that a more efficient encoding method is found. Thus, we predict that the tendency to move from *item-bound generalization* to *category-based generalization* is driven by an inflow of *noise,* which calls for a more efficient form of encoding, i.e. with less loss of information.

To this end, we add background *noise* while exposing the participants to the XXY grammar. The aim of this study is to disentangle the effect of adding background *noise* from the effect of overloading working memory with additional tasks. Although limited memory capacity has been proposed to promote generalization – under the classical *Less-is-More hypothesis* (Newport, 1990; 2016), previous studies found no effect of overloading working memory with additional tasks on generalization (Perfors, 2012). According to our entropy model, we hypothesize that *noise* drives generalization, rather than additional tasks that overload working memory. In order to test this prediction, we expose adults to the lowest entropy version of our XXY artificial grammar

(Radulescu et al., 2019), while playing random digits and beeps in the background to create a noisy environment. In one condition learners have an additional task besides listening to the XXY language, that is to pay attention and remember specific digits from the *noise* material (Dual-Task condition), while participants in another condition are not given any additional task on the background *noise* material (Distractor condition).

In addition to probing the direct effect of the *noise* variable of the *channel capacity* on rule induction, we also measure and control for the individual differences in relevant cognitive capacities: explicit/implicit memory capacity and a domain-general pattern-recognition capacity, which draws on working memory resources.

Chapter 7 lays the foundation of a new theoretical framework for rule induction by sketching an innovative research direction towards a thermodynamic theory of rule induction, which will complement the information-theoretic entropy model proposed in this dissertation. A comprehensive theory of rule induction requires the formulation of biologically plausible mechanisms in accord with the laws of biophysics and with evidence from neuroscience. Entropy-related concepts from information theory are ultimately linked to the same concepts in biophysics. This dissertation proposes an entropy model for rule induction whose main hypotheses posit that rule induction results from the brain's sensitivity to changes in information entropy interacting with channel capacity. But *why* is the brain sensitive to information entropy? And *how* does rule induction emerge? Information entropy is to a large extent a reflection of thermodynamic entropy (Karnani, Pääkkönen & Annila, 2009; Le Bellac, Mortessagne, & Batrouni, 2004; Sethna, 2006). Recent studies in biosciences converge on a thermodynamics view of the brain as an open system operating under the rule of the laws of physics, focusing on the laws of thermodynamics (Annila, 2016a, 2016b; Collell & Fauquet, 2015; DeCastro, 2013; Varpula, Annila, & Beck, 2013).

Chapter 7 proposes and sketches the first joint information-theoretic and thermodynamic model of rule induction. Specifically, this new perspective suggests that the 2nd law of thermodynamics can answer the question *why* rule induction happens, while the constructal law of thermodynamics can answer the question *how* rule induction happens.

In its first entropy-related formulation, the 2nd law of thermodynamics states: as heat always flows from hot to cold, entropy always increases (Feynman, Leighton, Sands, & Hafner, 1965). In modern phrasing, the 2nd law of thermodynamics states that spontaneously, energy always goes from being concentrated to being dispersed (Annila & Beverstock, 2016). Recent research in the physics of life forms supports the idea that the 2nd law of thermodynamics acts as a natural selection criterion that chooses organisms and mechanisms that are better at taking in and dispersing energy in the least time, in order to increase entropy (Annila & Annila, 2008; Avery, 2012). In other words, the principle of increasing entropy of the 2nd law equals the imperative to consume free energy. Free energy is the energy that can be used to produce useful work, unlike entropy (Schrödinger, 1944). In accord with the 2nd law of thermodynamics, we propose that rule induction is a natural result of the tendency of the brain's neural networks (and our cognitive system, consequently) to consume free energy (in the form of information), in the least time possible.

Proposed in the late 20th century, the constructal law is another first principle of thermodynamics (like the 2nd law), which is argued to account for the evolution of structure of all inanimate and animate systems in nature (Bejan, 1996; 2012). The constructal law states that every flow system evolves towards a particular structure that facilitates the flow of energy. A flow system is defined as everything that moves, animate or inanimate, i.e. a current or a stream originating from a point and moving to other points. We hypothesize that rule induction, just like any flow system in nature, has evolved for the purpose of facilitating faster and better flow (or transmission) of information. We propose that the constructal law predicts a particular design of the neural networks, and, as a reflection, of the cognitive system, which generates the hierarchical structure of rule induction (items and categories of items). This structure facilitates efficient information transmission, as a flow of energy.

Chapter 1

Patterns bit by bit. An Entropy Model for Rule Induction

Radulescu, S., Wijnen, F., and Avrutin, S.[1]

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 investigates what drives the inductive leap from memorizing specific items and statistical regularities to extracting abstract rules. We propose an innovative entropy model that offers one consistent information-theoretic account for both learning the statistical regularities in the input and generalizing to novel input. The model predicts that rule induction is an encoding mechanism

[1] This chapter is a slightly modified version of a published paper:
Radulescu, S., Wijnen, F., & Avrutin, S. (2019). Patterns bit by bit. An entropy model for rule induction. *Language Learning and Development.* Advance online publication. https://doi.org/10.1080/15475441.2019.1695620
OR
Radulescu, S., Wijnen, F., & Avrutin, S. (2020) Patterns Bit by Bit.
An Entropy Model for Rule Induction, *Language Learning and Development, 16:2*, 109-140, DOI: 10.1080/15475441.2019.1695620

The entropy model and parts of the findings were presented at the following conferences:
Architectures and Mechanisms for Language Processing : Edinburgh, UK, 2014
Architectures and Mechanisms for Language Processing : Valletta, Malta, 2015
CUNY Conference on Human Sentence Processing - University of Florida, Gainesville, USA, 2016
Statistical Learning Conference - Bilbao, Spain, 2017
Architectures and Mechanisms for Language Processing - Lancaster, UK, 2017
Architectures and Mechanisms for Language Processing - Berlin, Germany, 2018

The first version of the entropy model and the results of the first experiment were described in Silvia Radulescu's 2014 Master's Thesis, which is available from the Utrecht University Online Repository:
https://dspace.library.uu.nl/handle/1874/294595

gradually driven as a natural automatic reaction by the brain's sensitivity to the input complexity (*input entropy*) interacting with the finite encoding capacity 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 when *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. From memorizing sequences like *Dad walked slowly* and *Mom talked nicely*, to learning generalizations of the type "add –*ed*" to express a past action, and to generalizing to abstract categories (Noun, Verb, Adverb), and inducing a general rule that the sequence Noun-Verb-Adverb is well-formed, learners take a qualitative step from encoding exemplars to forming abstract categories and acquiring relations between them. This paper addresses this qualitative step from items to categories.

Following previous proposals in the literature (Gómez & Gerken, 2000), we will distinguish between two types of rule induction: *item-bound generalizations* and *category-based generalizations*. An *item-bound generalization* is a relation between perceptual features[2] 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 thoroughly understand language acquisition is how learners converge on these higher-order *category-based generalizations*.

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

1.2 Statistical learning vs. algebraic rules

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

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

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

1. Are these forms of encoding outcomes of two separate mechanisms, with *statistical learning* underlying *item-bound generalizations*, and *abstract rule learning* accounting for the higher-order *category-based generalizations*?

2. Or, are these forms of encoding two different outcomes of the same mechanism?

    a. If they are outcomes of the same mechanism, are the two types of generalizations stages of a phased mechanism that *gradually* transitions from a lower-level item-bound generalization to a higher-order abstract one?

    b. Or is it a mechanism that switches *abruptly* from one form of encoding to the other?

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

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

A subsequent study by Gerken (2010) may help finding answers. In this study, she exposed 9-month-olds to the same "ends in *di*" condition as in Gerken (2006), but – crucially – added three strings ending in "*je/we/li*" at the end of the familiarization. The participants subsequently made the broader AAB generalization. The author hypothesizes that the factor driving generalization is not the mere number of examples, but the logical structure of the input. She proposes that infants entertain incremental learning models (by updating their hypothesis in real time), and that they use rational decision criteria, in a process that resembles Bayesian learning. But we ask: would they make a broader generalization also if these 3 'divergent' strings were presented at the beginning of the 2-minute familiarization? Would infants not 'forget' those 3 strings, and rather update their model based on the more strongly evidenced and recent "end in *di"* input? As Gerken (2010) did not include this control condition, the study cannot decisively show that infants are incremental and "rational" learners, as there is no online measure or intermediate checkpoint into their models before and after each batch of stimuli. Nonetheless, it clearly shows that little evidence and variability is needed for them to move to a broader generalization. However, surprisingly, the results of Gerken, Dawson, Chatila, and Tenenbaum (2015) suggest that variability is not needed. An input consisting of a single item ("*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 (Gerken & Bollt, 2008; Gómez, 2002) show that input variability is a strong factor driving generalization. However, it seems that it is not mere variability that is critical, but a specific pattern of variable input. How can this specific pattern be captured and defined by incorporating all variables?

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 familiarized with nonsense strings having the underlying structure: *(Q)AXB (R)*[3], in order to probe whether they can generalize *X* as a category, rather than just memorize the exact strings. Participants heard different subsets of strings from this grammar, which displayed different combinations of items. In the test phase, participants were tested on the withheld (novel) grammatical strings, as well as on ungrammatical strings (*AXA* or *BXB* strings). In our terminology, participants' ability to recognize the novel strings as grammatical implies that they made the correct *category-based generalization* (i.e. *AXB*)*.* Reeder 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

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

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 hypothesizes that *rule induction is an encoding mechanism gradually driven as a natural automatic reaction by the brain's sensitivity to the input complexity (entropy) interacting with the finite encoding capacity of the human brain* (*channel capacity*). Thus, our model is based on the following tenets:

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

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

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

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

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

1. For a given set of *n* items from the input, entropy (H) is zero, if the probability of one item is 1 and the probabilities of all the other items are zero. Intuitively, this is a set with the lowest complexity, and *uncertainty.* In psychological terms, an event with only one outcome with a maximum probability of occurrence is totally predictable, i.e. the amount of surprise when that outcome occurs is zero.
2. For a given set of *n* items, the entropy is maximal if the distribution of the items' probabilities is uniform, i.e. when all the probabilities are equal (for example, for *n=4* and each item has *p=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.
4. Any change to render the probabilities of the items unequal (i.e. some items are more probable than others) causes a decrease in entropy.

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

C*hannel capacity* (C) describes the amount of entropy that can be sent through the channel per unit of time (Shannon, 1948). If H < C, information can be sent through the channel at the channel rate (*C)* with an arbitrarily small frequency of errors (equivocations) by using a proper encoding method. If H > C, it is possible to find an encoding method to transmit the signal over the channel, but the rate of transmission can never be higher than C. *Channel capacity* is employed here to model the finite encoding capacity of the information encoding system. Intuitively, the capacity to encode specific items and relations between them is finite. Thus, depending on the degree of input complexity and the finite encoding capacity (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 finite encoding capacity of the brain. The encoding mechanism moves gradually from an *item-bound* to a *category-based generalization* as a function of increasing *input complexity* (entropy), as follows:

a. If the input entropy is low – that is below or matches the *channel capacity*, then the input can be encoded using an encoding method that matches the input statistical structure, i.e. the probability distribution of the specific items in the input. Thus, the items with their specificity defined by their uniquely identifying features (acoustic, phonological, phonotactic, prosodic, distributional, etc.) and their specific probability distribution can be transmitted through the channel (i.e. encoded) at the default channel rate (i.e. amount of entropy per unit of time) and stored by *item-bound encoding* (i.e. probability matching to the input).
   Examples of *item-bound encoding* would include rules like "*ends in di",* or rules specifying what specific items would follow each other (e.g. *ba* or *ge* follows *wo*).

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

2. An increase of *channel capacity*, (e.g. resulting from growth/development), reduces the need, and thus the tendency to move to a higher-order *category-based* form of encoding. Therefore, if infants and adults are exposed to the same input entropy, adults will have a lower tendency to make a *category-based generalization* than infants, because adults' *channel capacity* is higher.
3. *Channel capacity* is used to model the finite encoding capacity of the human mind. We hypothesize that it is modulated by (unintentional) incidental memory capacity, attention and a general pattern-recognition capacity.

Therefore, the model hypothesizes that there is a *gradient of generalization*, in line with previous suggestions (Aslin & Newport, 2014), but it refines and extends this proposal, by further explaining how and why this gradual process happens. Sensitivity to entropy means a sensitivity to a specific pattern of variability in the input given by the degree of similarity/dissimilarity between items and their features and also their probability distribution, which assigns significance to specific items and their features. The more differences are encoded between specific items (i.e. many different specific features encoded for each item – measured in bits of information), the higher the degree of specificity of the encoding (i.e. *item-bound* specificity). Conversely, since the *channel capacity* places an upper bound on the number of bits encoded per unit of time, a reduction– *"gradual forgetting"* – of the encoded differences highlights more similarities, hence the lower the degree of specificity and the higher the degree of generality. Entropy captures this dynamics of specificity vs generality, and quantifies it in bits of information. Thus, a gradient of specificity/generality on a continuum from *item-bound* to *category-based encodings* can be envisaged in terms of less or more bits of information encoded in the representation.[4]

2.3. Application of the Model to AGL

Given that entropy is defined as a property of a variable[5], the input must be organizable in variables that can take certain values. In artificial grammar studies using patterns like AAB, AXB, each position of the patterns creates a variable (a category of items), whose possible values are the specific items: for example, variable A in a study on learning an AAB pattern (*le_le_di*) is filled by *le, wi, ji, de,* etc. Each category of bigrams and trigrams creates a variable, whose possible values are the specific bigrams and trigrams: for example, *lele* is a value of the AA category of bigrams, *ledi* is a possible value of the AB category of bigrams, while *wiwije* is one of the values taken by the AAB category of trigrams. Similarly, in finite-state grammars, the strings generated by the grammar can be segmented in groups of bigrams and trigrams, which can be construed as variables in a similar way. Given this set of variables, we can calculate the entropy of the familiarization input.

For an entropy model to be relevant for the encoding mechanism under scrutiny here, evidence is needed that learners acquire knowledge about categories of items that can be construed as variables: there is extensive evidence that grammaticality judgments in artificial grammar learning are shaped by knowledge acquired about bigrams and trigrams (Knowlton and Squire, 1994; Perruchet and Pacteau, 1990). Studies also showed that

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

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

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 familiarization 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 strings[6]. A lower entropy of a sequence of items (given by high probability bi-/trigrams and a low number of items) triggers a higher tendency to endorse it as possible in the familiarization language. Pothos's conclusions are in line with one of the predictions of our entropy model: a low entropy of the set of items enables *item-bound generalizations* (rules about which items follow each other).

3. A Unified Account for Previous Studies. A Brief Proof of Concept.

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

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

parsing other sets of bigrams/trigrams, we deem an average of bigram entropies and an average of trigram entropies to be the relevant measure. Also, based on the results reported by Pothos (2010) an average bigram/trigram entropy seems to be a better predictor for performance than the sum of all bigram/trigram entropies.

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

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

Table 1. Entropy values of the input in the Diagonal 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 insufficiently constrained and cannot account for this unchanged tendency in generalization. Consequently, their results are unexplained under their hypothesis. Just as Gerken (2010) suggested, it is not the mere number of exemplars that has an effect on generalization, but a specific pattern of variability. As we show in Table 3, this pattern of variability can be captured by input entropy. Even though the input was reduced to half the number of exemplars, the total entropy was only slightly reduced, which explains why learners' tendency to generalize remained almost the same. The entropy values of the set of stimuli used in Experiment 3 were significantly reduced as compared to the first two experiments, which can explain learners' lower likelihood to generalize the categories. The effect of increased exposure to the same stimulus set in the fourth experiment cannot be explained by the authors' hypothesis, as the input displayed the same statistical properties as in Experiment 3, but the tendency to generalize was significantly reduced. We would argue that increased exposure leads to stronger memory traces of the items, which allows for *item-bound generalization,* hence to a suppression of category-based generalization, which is in line with the predictions of our entropy model. The entropy values for Experiment 5 series (from A to D) are slightly higher than those for Experiment 1 – 4, respectively, which explains the slightly higher tendencies to generalize.

Column condition

[A	A	B]
le	le	di
wi	wi	di
ji	ji	di
de	de	di

Entropy values

$H[bA] = - [(p(le)*\log_2 p(le)) + (p(wi)*\log_2 p(wi)) + (p(ji)*\log_2 p(ji)) + (p(de)*\log_2 p(de))= 2$

$\mathbf{H[Be] = - [p(di)*\log_2 p(di)]= 0}$

$H[AA] = - [(p(lele)*\log_2 p(lele)) + (p(wiwi)*\log_2 p(wiwi)) + (p(jiji)*\log_2 p(jiji)) + (p(dede)*\log_2 p(dede))= 2$

$H[AB] = - [(p(ledi)*\log_2 p(ledi)) + (p(widi)*\log_2 p(widi)) + (p(jidi)*\log_2 p(jidi)) + (p(dedi)*\log_2 p(dedi))= 2$

$H[AAB] = - [(p(leledi)*\log_2 p(leledi)) + (p(wiwidi)*\log_2 p(wiwidi)) + (p(jijidi)*\log_2 p(jijidi)) + (p(dededi)*\log_2 p(dededi))]= 2$

H[bigram] = 1.5 H[trigram] = 2

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

In conclusion, our entropy model accounts for all the findings of these experiments and gives a complete and unifying picture of rule induction by capturing the specific pattern of input variability (*entropy*) interacting with exposure time (which affects working memory and therefore modulates *channel capacity*[7]). 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.*

Experiment_1	**Experiment_2**	**Experiment_3**	**Experiment_4**

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**

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)

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,* without specifically measuring variations in channel capacity (i.e. individual biological/psychological capacities), which were assumed to be roughly

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

insignificant since we tested participants of similar age and backgrounds. The following hypothesis was tested:

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

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

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

To test these predictions, we designed several versions of the same artificial grammar (3-syllable XXY structure[8]) in order to expose participants to different input entropies in three groups: high, medium and low entropy. An ensuing test phase presented participants with a grammaticality judgement task, where they were asked a yes/no question to indicate if they accepted the test strings as possible in the familiarization language. The test included four types of test strings that were designed to test each type of rule induction, as presented below.

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

New-syllable XYZ (XYZ structure with new syllables) **– correct answer: no - reject** – this is the complementary test case, which is intended to check learning of the familiar strings and string pattern. It is designed to back up and complement results for the familiar-syllable XXY strings as follows: if the forms of encoding – either ITEM[9] or CATEG – trigger acceptance of familiar XXY strings, then they should trigger rejection of the structurally and item non-compliant test cases (new XYZ). Thus, all groups are expected to reject this test type, with no between-group difference. If these strings are not consistently rejected, the interpretation of the results for familiar XXY cannot be valid.

New-syllable XXY (XXY structure with new syllables) **– correct answer: yes - accept** – this is a test case that is intended to be the TARGET test

[8] An XXY pattern describes strings consisting of two identical syllables (XX) followed by another different syllable (Y): e.g. *xoxofi* ; *pypydy*

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

string type to check generalization of rule to novel strings (*CATEG*). The number of correct answers is expected to be a function of entropy condition: the highest number of acceptances is expected in the high entropy group, followed by the medium, and the low entropy.

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

Familiar-syllable XYZ (XYZ structure with familiar syllables[10]) – **correct answer: no - reject** – this is the complementary test case to the New-syllable XXY strings: if New-syllable XXY strings are accepted in a different proportion by the three groups due to hypothesized differences in types of encoding developed, then Familiar-syllable XYZ strings should also be treated differently across groups. We expect results for this test type to capture the two types of encoding competing against each other, because it is likely that the memory trace of familiar syllables drives acceptance of these ungrammatical strings with familiar syllables. Hence differences in performance are expected across groups, depending on the extent to which ITEM and CATEG are developed, i.e. to the *gradient of generalization*: the low entropy group is expected to yield the highest proportion of correct rejections, as (per hypothesis) they encoded the strings as frozen item-bound generalizations, which highlight clear mismatches between familiar and non-compliant combinations of specific items. In the high entropy group, *category-based generalization* will be predominant, and thus XYZ strings will be rejected for being inconsistent with the XXY pattern. The medium entropy group is expected to yield the lowest percentage of correct rejections, because it is likely that the memory traces of the individual familiar

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

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. Familiarization stimuli

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

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

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

5.2. Entropy values of familiarization conditions

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

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

Low Entropy	Medium Entropy	High Entropy

H[bX]=H[6]= -Σ[0.167*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
H[total] = $\frac{H[bigram]+H[trigram]}{2}$ = 3.5
	H[bX]=H[12]= -Σ[0.083*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] = 4.25
	H[total] = 4
		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
		H[total] = 4.58

Table 4. Entropy values for Experiment 1

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

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

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

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 High Entropy, Medium Entropy and Low Entropy conditions were compared in a Generalized

Linear Mixed Model, with Accuracy (correct acceptance/rejection) as dependent variable and Entropy condition, Test String Type x Entropy condition interaction, Test phase x Entropy condition interaction as fixed factors, and Subject and Trial as random factors. An alpha level of *.05* was used for all statistical tests. We started fitting the data from the intercept-only model and added the random and fixed factors one by one. We report here the best fitting model, both in terms of model's accuracy in predicting the observed data, and in terms of AICc (Akaike Information Criterion Corrected). There was a statistically significant Test String Type x Entropy condition interaction (F(9, 679) = 6.363, *p* < .001). There was no statistically significant main effect of Entropy condition (F(2, 679) = 0.401, *p* = *.67*). Results indicated a non-significant trend in the predicted direction for Test phase x Entropy condition interaction (F(9, 679) = 1.243, *p* = *.26*).

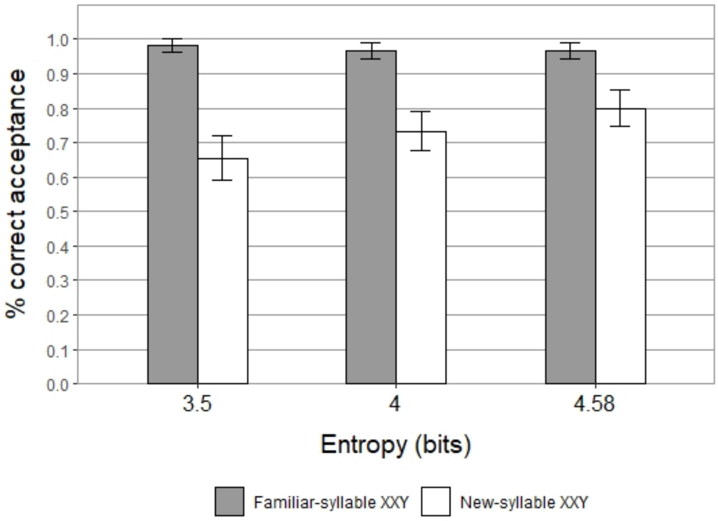


Fig. 1. Percentage of correct acceptance for Familiar-syllable XXY & New-syllable XXY. Error bars show standard error of the mean. Experiment 1.

Fig. 1 presents the mean rate of acceptance (percentage of acceptances per group) across conditions for Familiar-syllable XXY and New-syllable XXY. The mean acceptance rate of New-syllable XXY in High Entropy was 80% (Mean = .80, SD = .403), in Medium Entropy was 73% (Mean = .73, SD = .446), and in Low Entropy was 65% (Mean = .65, SD = .480). One-sample Wilcoxon Signed-Rank tests indicated a statistically significant above-chance mean acceptance for New-syllable XXY in High Entropy (*Z = 4.648, SE = 118.12, p < .001;* Cohen's effect size *d = 0.6*), in Medium Entropy *(Z = 3.615, SE = 118.12, p < .001;* Cohen's effect size *d = 0.47),* and in Low Entropy *(Z = 2.292, SE = 103.82, p = .022,* Cohen's effect size *d = 0.31*). In High Entropy there was a significant difference between acceptance of Familiar-syllable XXY and acceptance of New-syllable XXY (M=.167, SD=.376; *t(*3) = 2.721, SE=0.853, *p* = .007); in Medium Entropy there was also a significant difference between performance on these tests (M=.233, SD=.427; *t(*3) = 3.454, SE=0.838, *p* = .001); and in Low Entropy the difference between performance on these tests was also significant (M=.327, SD=.511; *t(*3) = 3.566, SE=1.158, *p* < .001). 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 High Entropy vs. Low Entropy were higher than the same values for High Entropy vs. Medium Entropy (d = 0.15, r = 0.07), and also higher than the same values for Low Entropy vs. Medium Entropy (d = 0.21, r = 0.1).

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

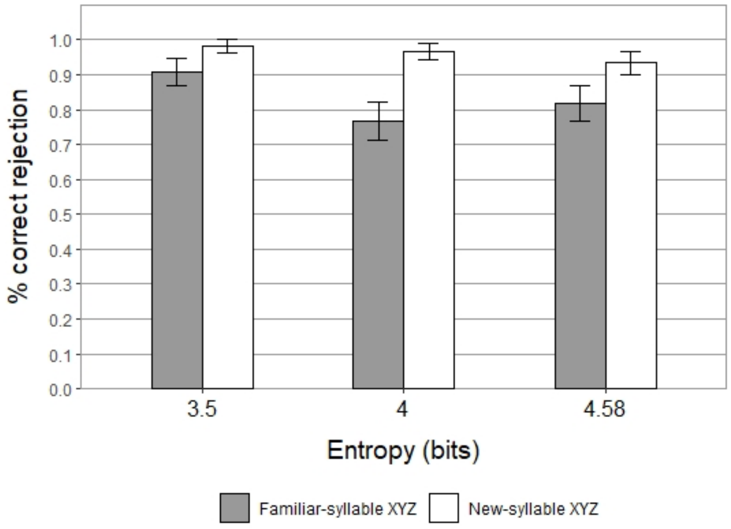


Fig. 2. Percentage of correct rejection for Familiar-syllable XYZ & New-syllable XYZ. Error bars show standard error of the mean. Experiment 1.

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. familiar XXY strings depending on the entropy group. This shows differences between groups in terms of how learners encode the XXY strings: if the participants do not make a clear distinction between a new XXY and a familiar XXY, we conclude that they formed a *category-based generalization* (XXY) which applies equally to both familiar and new XXY strings. Thus, a smaller difference between the means of acceptance of these test types shows a higher tendency to make *category-based generalizations*. The results showed that in the high entropy group this difference is smaller than in the medium entropy one, which is smaller than in the low entropy group. Hence these results indicate that learners exposed to higher *input complexity* had a higher tendency to make *category-based generalizations* and to generalize to novel strings displaying the underlying XXY pattern, which is in line with the predictions of our entropy model.

The rate of correct rejection for XYZ strings with familiar syllables is very high in the low entropy group, although the rate of acceptance for new XXY strings is rather low (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 familiar syllables, but only as high as the acceptance of new XXY strings. This indicates that CATEG is strong enough to drive rejection of the XYZ strings. As predicted, the medium entropy group yielded the lowest performance of all groups. The interpretation is that increased *input complexity* prevents a strong memory trace of the entire strings, and thus ITEM cannot support a consistent and confident rejection of the XYZ strings. At the same time, CATEG is not strongly developed to consistently reject the incorrect XYZ pattern. To sum up, the results showed a roughly U-shaped performance on XYZ with familiar syllables, as a function of increased input entropy. Similar tendencies 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.[11]

The results showed that the decreasing trend of the rejection of familiar-syllable XYZ changes into an increasing trend roughly at the same entropy level where it meets the increasing trend of acceptance of new XXY. We hypothesize that the lowest point of the U-shaped trend of the rejection of familiar-syllable XYZ is the intersection point of the decreasing trend of XYZ and the increasing trend of XXY. The calculated intersection point of the two trends – *y(New-syllable XXY) = y(Familiar-syllable XYZ)* – is *H = 4.2* (*y = 0.72)*, which allows the prediction that the rate of rejection of Familiar-syllable XYZ decreases until 72%, if the *input complexity* is H=4.2 bits. This value is predicted to be the point where the

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

decreasing trend for Familiar-syllable XYZ reaches its minimum and changes into an increasing function, given that CATEG outperforms ITEM. This point is hypothesized to roughly mark the excess limit of the *channel capacity*.

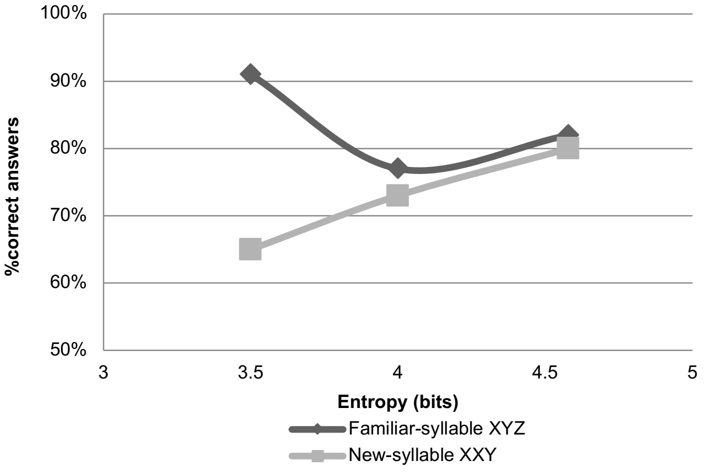


Fig. 3. Percentage of correct acceptance of New-syllable XXY and correct rejection of Familiar-syllable XYZ plotted against input entropy. Experiment 1

A subsequent re-thinking of the XYZ strings with familiar syllables raised the question that these strings should have had an X_1X_2Y pattern (X_1 is different from X_2), to ensure that the reason for the rejection of these strings does not involve the inconsistency of using X-syllables in the last position of the strings, or Y-syllables in the first or second position of the string. Only two out of five Familiar-syllable XYZ strings did not have an X_1X_2Y pattern. However, this confound would have helped rejection of these strings more in the low entropy group, where it was easier to remember the specific familiar X-syllables and Y-syllables. An ANOVA with familiarization group (High Entropy, Medium Entropy and Low Entropy as between-subjects variable and test item (X_1X_2Y vs. non-X_1X_2Y) as within-subjects variable revealed no statistically significant difference between the rejection rate of X_1X_2Y strings and the rejection rate of the non-X_1X_2Y strings in any of the conditions (High Entropy: Mean[X_1X_2Y] = .81, Mean[non-X_1X_2Y] = .83, F(1,58) = .072, *p = .79*; Medium Entropy: Mean[X_1X_2Y] = .79, Mean[non-X_1X_2Y] = .73, F(1,58) = .293, *p = .59*; Low Entropy: Mean[X_1X_2Y] = .91, Mean[non-X_1X_2Y] = .91, F(1,53) < .001, *p = 1.00*. Therefore, such a confound is highly unlikely to explain the results.

We designed intermediate tests to investigate the learning process as an interaction between input entropy and exposure time. On the one hand, we predicted that longer exposure to the familiarization items would strengthen the memory trace of the specific items, and thus it would make it easier to encode the specific syllables/strings. Thus, the tendency to make *category-based* *generalizations* will decrease as a function of increasing exposure time, as was shown in Reeder et al (2013). On the other hand, a high input entropy would make remembering the specific items more difficult than a medium entropy and a low entropy. Thus, an interaction between input entropy and exposure time was predicted to show the following results: the acceptance of new XXY strings across the intermediate tests through the final test is expected to decrease in all entropy groups due to exposure time. But at a different rate, depending on the input entropy, as follows: the percentage of acceptance of new XXY strings should have a slowly decreasing trend in high entropy (because the more complex input prevents forming memory trace of specific items and strings), a slightly steeper decreasing trend in medium entropy, and an even steeper decreasing trend in low entropy (because the more repetitive input allows remembering of specific items and strings). Although the results did not reach statistical significance, the trends match the predictions: in low entropy group the performance curve decreases slightly steeper than in the medium entropy, and steeper than in the high entropy one. Further research would need to be conducted with larger samples and longer exposure time to further investigate the generalization curve as an interaction between input entropy and exposure time.

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 Low Entropy (Experiment 2) condition we chose a lower entropy value than for Low Entropy (Experiment 1) (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 Medium Entropy (Experiment 2) 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 familiar syllables will decrease as compared to the performance for Medium Entropy (Experiment 1) (for *H=4 bits* the performance was 77%): at *H=4.2 bits* the mean performance predicted is 72%. For the High Entropy (Experiment 2) condition we chose a higher entropy (4.8 bits - 1 x 28 Xs / 1 x 28 Ys) than High Entropy (Experiment 1) in order to test if the tendency to abstract away from the specific input increases further or it stabilizes at a certain ceiling. The prediction is that at a certain degree of entropy the tendency to generalize will stabilize at a certain ceiling regardless of how much the entropy increases, due to the finite *channel capacity*, i.e. there will be no further increase in the tendency towards *category-based encoding*.

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. Familiarization stimuli

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

8.2. Entropy values of familiarization conditions

The Shannon entropy formula and the entropy calculations were applied in the same manner as for Experiment 1 to obtain other three different values for *input complexity*, as follows:
1. **Low Entropy:** 7 X-syllables and 7 Y-syllables (with each syllable used 4 times in each familiarization phase). To generate the XXY strings for the Low Entropy condition, the 7 XX pairs were concatenated with the 7 Y-syllables to obtain 7 strings, which were repeated 4 times to obtain 28 strings which were used in all familiarization phases. The same procedure was applied to the other conditions. All three familiarization phases had the same entropy values: the average bigram entropy (H[bigram]) was 2.8, the average trigram entropy (H[trigram]) was 2.8, and the total average entropy (H[total]) was 2.8 (see Table 5 for complete entropy calculations).
2. **Medium Entropy**: 14 X-syllables and 14 Y-syllables (7 different X-syllables and 7 different Y-syllables were added to those used for Low Entropy with each syllable used 2 times. All three familiarization phases had the same entropy values: the average bigram entropy (H[bigram]) was 4.05, the average trigram entropy (H[trigram]) was 4.46, and the total average entropy (H[total]) was 4.25.
3. **High Entropy**: 28 X-syllables and 28 Y-syllables (14 X-syllables and 14 Y-syllables were added to those used for Medium Entropy with each syllable used one time. All three familiarization phases had the same entropy values: the average bigram entropy (H[bigram]) was 4.8), the average trigram entropy (H[trigram]) was 4.8, and the total average entropy (H[total]) was 4.8.

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

8.3. Procedure

The procedure was the same as for Experiment 1.

Low Entropy	Medium Entropy	High Entropy

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

Table 5. Entropy values for Experiment 2

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):

Familiar-syllable XXY– correct answer: yes - accept

New-syllable X_1X_2Y (three different new syllables) **– correct answer: no - reject**

New-syllable XXY – correct answer: yes - accept

Familiar-syllable X_1X_2Y (three different familiar syllables) **– correct answer: no - reject**

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

9. Experiment 2: Results

In order to test the effect of *input complexity* on the process of generalizing, the High Entropy, Medium Entropy and Low Entropy conditions were compared in a Generalized Linear Mixed Model, with Accuracy (correct acceptance/rejection) as dependent variable and Entropy condition, Test String Type x Entropy condition interaction, Test phase x Entropy condition interaction as fixed factors, and Subject and Trial as 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* < .001). There was no statistically significant main effect of Entropy condition (F(2, 699) = 0.260, *p = .77*). Results indicated a non-statistically significant trend in the predicted direction for Test phase x Entropy Group interaction (F(9, 699) = 1.163, *p = .32*).

Fig. 4 shows the mean acceptance rates across conditions for Familiar-syllable XXY and New-syllable XXY. The mean rate of acceptance for New-syllable XXY in High Entropy was 80% (Mean = .80, SD = .403), for Medium Entropy was 77% (Mean = .77, SD = .427), and for Low Entropy was 57% (Mean = .57, SD = .5). One-sample Wilcoxon Signed-Rank tests indicated a statistically significant above-chance mean acceptance for New-syllable XXY in High Entropy (*Z = 4.648, SE = 118.12, p* < .001*;* Cohen's *d = 0.6*) and in Medium Entropy *(Z = 4.131, SE = 118.12, p* < .001*; d = 0.53),* but in Low Entropy the mean acceptance was not significantly above chance *(Z = 1.033, SE = 118.12, p = .3, d = 0.13*). In High Entropy there was a significant difference between acceptance of Familiar-syllable XXY and acceptance of New-syllable XXY (M=.167, SD=.376; *t(*3) = 2.161, SE=0.643, *p* = .031); in Medium Entropy there was also a significant difference between performance on these tests (M=.233, SD=.427; *t*(3) = 2.542, SE=0.624, *p* = .011); and in Low Entropy the difference between performance on these tests was also significant (M=.327, SD=.511; *t(*3) = 4.335, SE=0.683, *p* < .001). Further, Cohen's d (d = 0.73) and the effect-size correlation (r = 0.34) for the difference between acceptance of Familiar-syllable XXY and acceptance of New-syllable XXY in High Entropy vs. Low Entropy were higher than the same values for High Entropy vs. Medium Entropy (d = 0.09, r = 0.04), and also higher than the same values for Low Entropy vs. Medium Entropy (d = 0.63, r = 0.3).

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

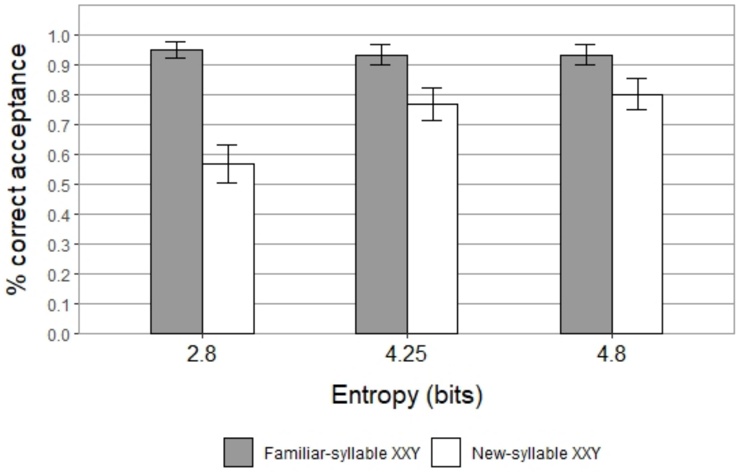


Fig. 4. Percentage of correct acceptance for Familiar-syllable XXY & New-syllable XXY. Error bars show standard error of the mean. Experiment 2.

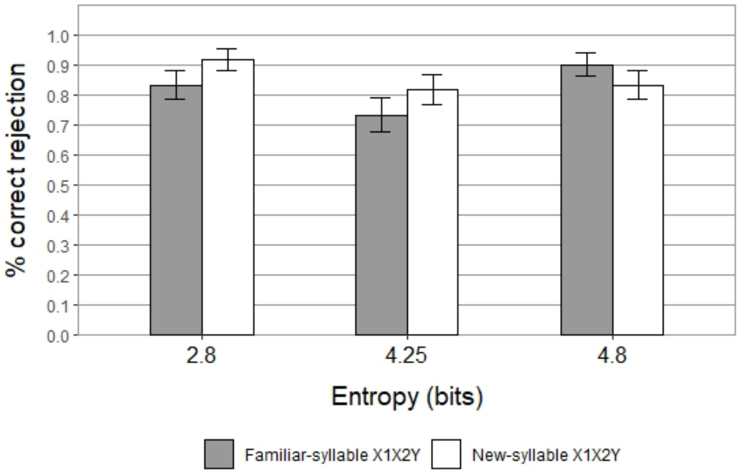


Fig. 5. Percentage of correct rejection for Familiar-syllable X1X2Y & New-syllable X1X2Y. Error bars show standard error of the mean. Experiment 2.

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 Experiment 1 and Experiment 2 were combined in an omnibus Generalized Linear Mixed Model, with Accuracy (correct acceptance/rejection) as dependent variable and Entropy condition, Test String Type x Entropy condition interaction, Test phase x Entropy condition interaction as fixed factors, and Subject and Trial as random factors. An alpha level of *.05* was used for all statistical tests. We started fitting the data from the intercept-only model and added the random and fixed factors one by one. We report here the best fitting model, both in terms of model's accuracy in predicting the observed data, and in terms of AICc (Akaike Information Criterion Corrected). There was a statistically significant Test String Type x Entropy condition interaction (F(18, 1,378) = 5.782, *p < .001*). There was no statistically significant main effect of Entropy condition (F(5, 1,378) = 1.165, *p = .32*), and no statistically significant Test phase X Entropy condition interaction (F(18, 1,378) = 1.150, *p = .29*).

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

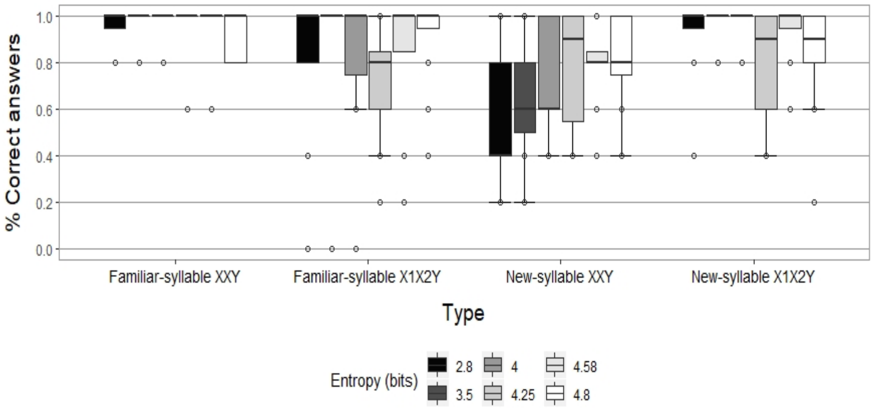


Fig. 6. On the X-axis, the four types of test items: Familiar-syllable XXY; Familiar-syllable X1X2Y; New-syllable XXY; New-syllable X1X2Y. On the Y-axis, the mean rate of correct answers: correct acceptance for XXY strings (with familiar or new syllables) and correct rejection for X1X2Y strings (with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

A simple linear regression was calculated (Fig. 7) to predict the rate of acceptance of New-syllable XXY based on the amount of input entropy. A significant regression equation was found (F(1,4)=243.54, p < .001), with an R^2 of .98. Input entropy was a significant predictor for the rate of acceptance of New-syllable XXY.

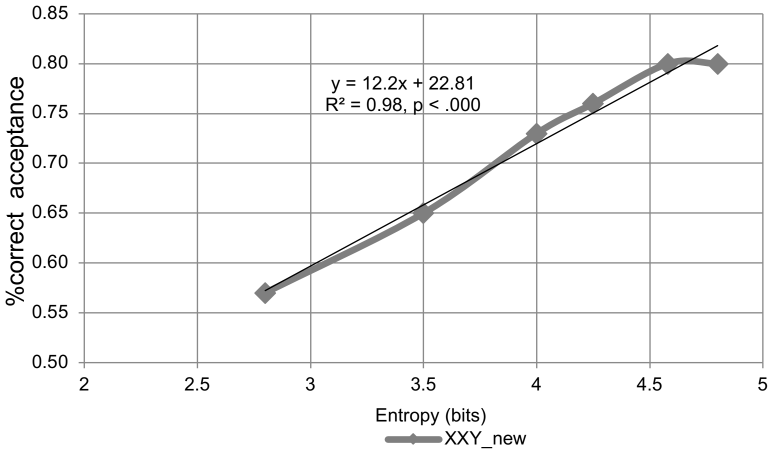


Fig. 7. Percentage of acceptance for New-syllable XXY as a function of input entropy. Experiments 1 & 2

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

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

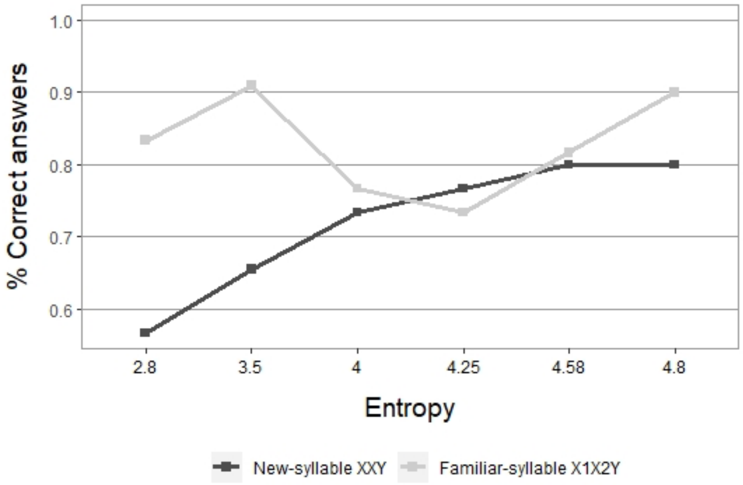


Fig. 8. Percentage of correct acceptance of New-syllable XXY and correct rejection of Familiar-syllable X1X2Y. Experiments 1 & 2.

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. 9) to predict uncertainty regarding the XXY structure of the input, based on the amount of input entropy. A significant logarithmic equation was found (F(1,4)=321.63, p < .001), with an R^2 of .98. As shown in Fig. 9, the uncertainty about structure 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*

(Endress & Bonatti, 2007; Marcus et al., 1999), while others proposed that *statistical learning* underlies both types of generalizations (Aslin & Newport, 2012; 2014; Frost & Monaghan, 2016; Perruchet & Pacton, 2006). Recent computational models suggest that learners might combine statistical learning and rule-based learning (Adriaans & Kager, 2010; Frank & Tenenbaum, 2011). However, these studies do not explain how the two mechanisms relate to each other, and it has remained unclear if and how two qualitatively different forms of encoding (*item-bound* and *category-based generalizations)* can arise from a single mechanism. Our model and the results of our experiments support the view put forth by Aslin and Newport (2012; 2014). These authors suggested that it is the (in)consistency of the distribution of contextual cues that triggers a narrow generalization (*item-bound generalization,* in our terminology) or a broader generalization (*category-based generalization*). However, they did not provide a precise description of the pattern of such (in)consistencies, and their hypothesis cannot answer the following questions: 1) What is the specific pattern of (in)consistencies and how much (in)consistency is needed to move from *item-bound* to *category-based generalization*? 2) What triggers this transition? 3) Why infants (children) and adults need different degrees of (in)consistency? Some studies pointed to memory constraints, under the *Less-is-More* hypothesis, but without clear evidence or explanation (Hudson Kam & Chang, 2009; Hudson Kam & Newport, 2005; 2009; Newport, 1990; Newport, 2016). 4) Why does increased exposure to the same distribution of (in)consistent cues reduce the tendency to make *category-based generalizations*?

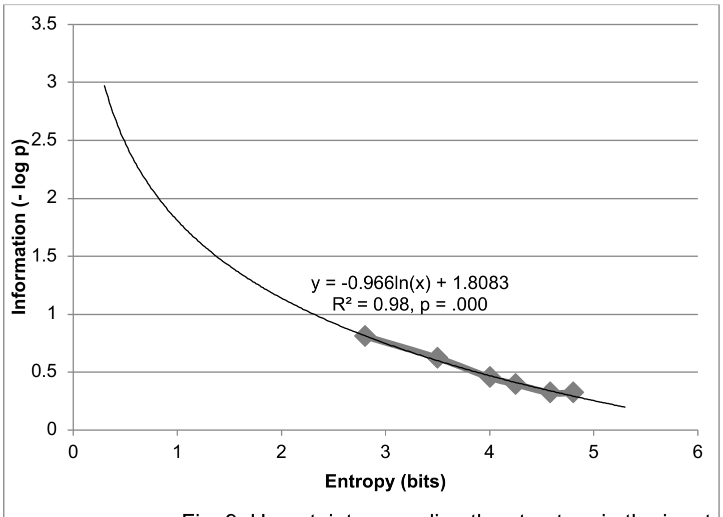


Fig. 9. Uncertainty regarding the structure in the input

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

Our model is in line with the general *Less-is-More* hypothesis, and it offers an extended and more refined formal approach to this hypothesis. Moreover, our model is in line with evidence from neurobiology (Frankland, Köhler, & Josselyn, 2013; Hardt, Nader, & Wang, 2013; Migues et. al, 2016; Richards & Frankland, 2017) and neural networks research (Hawkins, 2004; Kumaran, Hassabis, & McClelland, 2016; MacKay, 2003) that converge on the findings and hypothesis that the memory system is designed to remember a certain degree of specificity (i.e. of entropy, in our terminology) in order to prevent underfitting (missing specific parameters to correctly capture the underlying structure of the data), but also to prevent overfitting to past data/events (inadequately remembering and representing noise as underlying structure). According to these hypotheses and findings, rather than being faithful in-detail representations of the past data/events, memories are models optimized for future data integration, i.e. for better generalization and prediction of future data/events, in order to allow for more flexibility and better adaptability to noisy environments. As a refined information-theoretic extension of the *Less-is-More* hypothesis and in accord with these current developments in neurobiology and neural networks research, our entropy model offers a basis for conceptualization and quantification of the specific pattern of variability (input entropy) that the brain is naturally sensitive to, and which drives in a gradual fashion the rule induction mechanism in order to prevent overfitting to the input and to allow for representations of novel future input. From an information-theoretic point of view, our model proposes *channel capacity* (amount of entropy processed per unit of time) to reflect and quantify this design feature of the memory system proposed in neurobiology and neural network research that naturally and automatically places a lower and an upper bound on the degree of specificity (quantified in bits of information) represented in the neural pathways when encoding information, i.e. creating memory representations as actively predictive models of novel data/events. *Channel capacity* adds into the rule induction "formula" the essential dimension of time, i.e. a rate of encoding the entropy in the environment, as a natural physical system that is sensitive to a time-dependent and noisy (= highly entropic) inflow of information (Radulescu, Murali, Wijnen, & Avrutin, *(*2021).

In two artificial grammar experiments we tested the model by investigating the effect of one factor of the model, namely *input entropy,* on rule induction. The findings strongly support the predictions of our entropy model, namely: *item-bound generalization* and *category-based generalization* are not independent outcomes of two qualitatively different mechanisms. Rather, they are outcomes of the same information encoding mechanism that *gradually* moves from a lower-level *item-bound* encoding to a higher-order abstract encoding (*category-based generalization)*, as triggered by the *input entropy:* the lower the *input entropy*, the higher the tendency 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, the following outstanding questions should be investigated.

What is the effect of *input entropy* on infant rule induction? Further investigation is needed in order to probe whether the same pattern of results found in adults is replicated in infants, i.e. infants' tendency 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 should be exposed to a lower range of entropy than adults. Previous research into infants' generalization mechanisms have already hinted at the significance of surprise (in our terminology, *entropy*) as a triggering factor for generalization (Gerken et al., 2015). However, the necessary amount and nature of input variability (or surprise) remains unclear: some studies show that at least three or four examples are needed for infants to generalize (Gerken, 2006; 2010; Gerken & Bollt, 2008; Peterson, 2011), but Gerken et al. (2015) claim that a single example suffices for generalization. Gerken et al. (2015) interpreted their results to support a Bayesian account of generalization, also suggested by Griffiths & Tenenbaum (2007): when an input is inconsistent with learners' prior model, hence surprising, learners seek a new hypothesis to accommodate the new (surprising) input. However, we think that these results raise concerns. Firstly, the authors used a very reduced exposure time (21 seconds) compared to previous studies – 2 minutes in Gerken (2006; 2010). Reduced exposure time is a crucial component in the mechanisms of rule induction, as noted in previous studies with adults (Reeder et al., 2013), and as explicitly predicted by the time-dependent *channel capacity* component of our entropy model. Secondly, the authors claim that generalization occurred from a single example, which is surprising compared to their prior model. Formally, learners' analysis encompasses also their prior model, not just the one example they were exposed to in the lab. And we think (although the authors do not take this into account) that infants' analysis and learning extend also over the very long test phase (much longer than the familiarization phase itself), which includes 12 test trials with added variability (four different examples). Considering these concerns, the conclusion that infants generalize only from a single example is not decisive, and thus further research is needed to capture the nature and specific pattern of entropy (i.e. surprise) that drives infant rule induction.

In this paper, we proposed an original implementation of entropy as a quantitative measure of input complexity to artificial grammar learning with adults, but testing our model of a *gradual* transition from *item-bound* to *category-based generalization* with infants will require a different approach to implementation, in terms of calculations of entropy which should be different for infants, given that their cognitive system is still under development, so their *channel capacity* is hypothesized to be reduced. Infants might be more sensitive to local statistical properties of the input, rather than the entire set of items, and they might update their memory representations in an incremental fashion, as suggested already by evidence found in infant research (Gerken, 2010; Gerken & Quam, 2017). Thus, indeed due to a lower encoding capacity (*channel capacity*), underpinned by more plasticity of their developing memory system, infants' learning system may not be sensitive to average of bigrams/trigrams over the entire set of stimuli, since their encoding "window" might be more locally tuned (lower channel capacity). Moreover, since infants' sensitivity to similarities vs differences might develop gradually in the first year of life, given evidence that a primitive similarity detector is in place from birth (Gervain et al., 2008) and a detector for differences might develop later around 6-7 month old, as suggested by our recent findings (Radulescu, Wijnen, Avrutin, & Gervain, 2021 – see Chapter 2). As hypothesized by our model, sensitivity to entropy encompasses both a sensitivity to similar (or identical) features, and also a sensitivity to differences, thus these should be developmentally available for the sensitivity to entropy to be fully fledged.

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

The results presented in this paper point to a ceiling effect of input entropy on rule induction, which is the result of the brain's finite encoding capacity, captured by the *channel capacity* factor in our model. We hypothesize that the encoding capacity varies according to individual differences in (unintentional) incidental memory and a general pattern-recognition capacity. We have already found evidence for a negative effect of incidental memorization and a positive effect of a visual pattern-recognition capacity on rule induction (Radulescu, Giannopoulou, Avrutin, &Wijnen 2021 – see Chapter 3).

Further research should be conducted to investigate the suitability and feasibility of entropy as a quantitative measure of input complexity and of learners' uncertainty (i.e. surprise) in rule induction, and also to assess the generalizability of this model to more complex non-repetition grammars. As suggested by previous studies (Endress, Dehaene-Lambertz & Mehler, 2007; Endress, Nespor, & Mehler, 2009) a low-level perceptual identity-detector ("repetition detector"), which is in place from birth (Gervain, Berent, & Werker, 2012; Gervain, Macagno, Cogoi, Peña, and Mehler, 2008), might aid learning of repetition-based grammars. Indeed, we assume that our entropy model is generalizable to all grammars and further investigations are needed to probe its implementation and feasibility. In a recent study on non-adjacent dependencies learning that extends and refines previous findings by Gómez (2002) we found that the mere set size of items was not the only factor to drive generalization, but it was the specific pattern of variability captured by input entropy, as predicted by our entropy model (Radulescu and Grama, 2021).

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). Moreover, entropy was shown to have an effect on lexical access in unimpaired adults as well as in elderly populations and individuals with non-fluent aphasia (Van Ewijk, 2013; Van Ewijk & Avrutin, 2016). Entropy also plays an important role in other cognitive mechanisms beyond language learning, for instance in decision-making (Tversky & Kahneman, 1992) and problem-solving (Stephen, Dixon & Isenhower, 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, & 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).

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 entropy* and the finite *channel capacity* that drives the *gradual* transition to an abstract-level generalization, this research fills in an important gap in the puzzle about the induction problem for language acquisition.

Appendix

Familiarization items. Experiment 1.

High Entropy			Medium Entropy			Low Entropy		
Phase 1	Phase 2	Phase 3	Phase 1	Phase 2	Phase 3	Phase 1	Phase 2	Phase 3
xoxoʃi	xoxoʋu	xoxoʋu	xoxoʃi	xoxokaː	xoxokeː	xoxoʃi	xoxoʃi	xoxokeː
pypydy	pypyʃøː	pypysaː	pypyʃi	pypykaː	pypysaː	pypyʃi	pypysaː	pypysaː
tøːtøːsaː	Tøːtøːby	tøːtøːby	tøːtøːdy	tøːtøːmy	tøːtøːsaː	tøːtøːʃi	tøːtøːdy	tøːtøːʃi
veːveːføː	veːveːmo	veːveːdaː	veːveːdy	veːveːmy	veːveːmy	veːveːʃi	veːveːʃi	veːveːʃi
ʋoʋo mo	ʋoʋofaː	ʋoʋokeː	ʋoʋosaː	ʋoʋoɣo	ʋoʋoɣo	ʋoʋo dy	ʋoʋo dy	ʋoʋosaː
lolokeː	Loloko	lolody	lolosaː	loloɣo	loloɣo	lolody	loloʃi	loloʃi
xuxuʃu	xuxu møː	xuxuʃi	xuxuføː	xuxuʃi	xuxuʃi	xoxody	xoxosaː	xoxody
høːhøːʋøː	høːhøːxi	høːhøːko	høːhøːføː	høːhøːʃi	høːhøːʃi	pypy dy	pypy dy	pypy dy
jyjyfi	jyjyzy	jyjyʃøː	jyjymo	jyjydy	jyjydy	tøːtøːsaː	tøːtøːsaː	tøːtøːdy
ninikaː	ninify	niniføː	ninimo	ninidy	niniføː	veːveːsaː	veːveːdy	veːveːdy
roromy	rorobo	roromo	rorokeː	rorosaː	roroføː	ʋoʋosaː	ʋoʋoʃi	ʋoʋoʃi
vyvyɣo	vyvyhy	vyvyfaː	vyvykeː	vyvysaː	vyvy mo	lolosaː	lolosaː	lolosaː
haːhaːʋu	haːhaːʃi	haːhaːmøː	xoxoʃu	xoxoføː	xoxoʋøː	xoxoføː	xoxoføː	xoxosaː
hihiʃøː	hihidy	hihizy	pypyʃu	pypyføː	pypyfi	pypyføː	pypykeː	pypykeː
jijiby	jijisaː	jijixi	tøːtøːʋøː	tøːtøːmo	tøːtøːfi	tøːtøːføː	tøːtøːmo	tøːtøːføː
jujudaː	jujuføː	jujufy	veːveːʋøː	veːveːmo	veːveːmo	veːveːføː	veːveːføː	veːveːføː
jøːjøːfaː	jøːjøːdaː	jøːjøːbo	ʋoʋofi	ʋoʋokeː	ʋoʋokeː	ʋoʋo mo	ʋoʋo mo	ʋoʋokeː
liliko	lilikeː	lilihy	lolofi	lolokeː	lolody	lolomo	loloføː	loloføː
lylymøː	lylyʃu	lylyʃu	xuxukaː	xuxuʃu	xuxuʃu	xoxo mo	xoxokeː	xoxo mo

nonoxi	nonoʋøː	nonoʋøː	høːhøːkaː	høːhøːʃu	høːhøːʃu	pypymo	pypymo	pypymo
nunuzy	nunufi	nunufi	jyjymy	jyjyʋøː	jyjyʋøː	tøːtøːkeː	tøːtøːkeː	tøːtøːmo
ryryfy	ryrykaː	ryrykaː	ninimy	niniʋøː	ninikaː	veːveːkeː	veːveːmo	veːveːmo
vivibo	vivimy	vivimy	roroɣo	rorofi	rorokaː	ʋoʋokeː	ʋoʋoføː	ʋoʋoføː
vøːvøːhy	vøːvøːɣo	vøːvøːɣo	vyvyɣo	vyvyfi	vyvymy	lolokeː	lolokeː	lolokeː

Test items. Experiment 1.

	Test 1	Test 2	Test 3	Final Test	
Familiar-syllable XXY	xoxoʃi	veːveːmo	ʋoʋokeː	pypysaː	lolody
New-syllable XYZ	dovaːsøː	røːluxeː	mitaːzu	fuseːbi	køːsodo
New-syllable XXY	pøːpøːdeː	totosy	vovofo	ʋaːʋaːzøː	xøːxøːki
Familiar-syllable XYZ	tøːdysaː	pyʋoføː	loxomo	veːdyʋo	tøːveːkeː

Familiarization items. Experiment 2.

High Entropy Phase 1/2/3	**Medium Entropy** Phase 1/2/3	**Low Entropy** Phase 1/2/3
keːkeːmy	keːkeːmy	keːkeːmy
jujuɣo	jujuɣo	jujuɣo
daːdaːli	daːdaːli	daːdaːli
pypyveː	pypyveː	pypyveː
tøːtøːrøː	tøːtøːrøː	tøːtøːrøː
hihisaː	hihisaː	hihisaː
fofoʃu	fofoʃu	fofoʃu
nonoʃøː	nonoʃøː	keːkeːmy
nunuvøː	nunuvøː	jujuɣo
kykyʋaː	kykyʋaː	daːdaːli
jøːjøːvi	jøːjøːvi	pypyveː
totomøː	totomøː	tøːtøːrøː
haːhaːvy	haːhaːvy	hihisaː
fyfyʃi	fyfyʃi	fofoʃu
dodoɣøː	daːdaːmøː	keːkeːmy
bybyro	pypyvy	jujuɣo

bibimo	tøːtøːʃi	daːdaːli
kikiɣu	hihimy	pypyveː
fifizy	fofoɣo	tøːtøːrøː
fufuʋøː	nonoli	hihisaː
høːhøːʋo	nunuveː	fofoʃu
kaːkaːzøː	kykyrøː	keːkeːmy
køːkøːlu	jøːjøːsaː	jujuɣo
boboɣeː	totoʃu	daːdaːli
deːdeːvaː	haːhaːʃøː	pypyveː
hyhysøː	fyfyvøː	tøːtøːrøː
faːfaːly	keːkeːʋaː	hihisaː
jyjyxi	jujuvi	fofoʃu

Test items. Experiment 2.

Test 1		Test 2	Test 3	Final Test	
Familiar-syllable XXY	daːdaːli	hihisaː	keːkeːmy	tøːtøːrøː	jujuɣo
New-syllable X1X2Y	poxaːru	runyni	xaːmisy	syniny	mininy
New-syllable XXY	dydytaː	zuzuvo	sosory	jijiføː	ʋuʋuseː
Familiar-syllable X1X2Y	judaːsaː	pytøːmy	keːfoveː	hidaːrøː	tøːpyɣo

Chapter 2

Same Processing Costs for Repetition and Non-Repetition Grammars in 6-month-olds: An fNIRS Study

Radulescu, S., Wijnen, F., Avrutin, S., and Gervain, J.[12]

Abstract

How does encoding of linguistic regularities such as repetition regularities (e.g. ABB "bu-ra-ra") develop in infancy? Previous studies showed that 7-month-olds can recognize and encode such repetition regularities as abstract rules (Marcus et al., 1999), but only when the input showed some variability (Gerken, 2006). However, the nature, the developmental trajectory and the neural correlates of these mechanisms remain still largely unexplained. In an fNIRS study, we tested whether and how 6-month-old infants process and encode repetition-based linguistic regularities (ABB) as compared to non-repetition controls (ABC, e.g. "bu-fa-zo"), and also the effect of input entropy on encoding these patterns. According to an entropy model we proposed for rule induction in adults (Radulescu et al., 2019), we hypothesized that input entropy would have a positive effect on rule learning, such that a higher input entropy would support better discrimination between ABB and ABC patterns. In a channel-by-channel analysis, we found significant activation compared to baseline for both the ABB and the ABC conditions. In the same analysis, we also found higher activation for ABC in High Entropy than ABC in Low Entropy in three channels, higher activation for ABB in High Entropy than ABB in Low Entropy in one channel, and also higher activation for ABC than ABB in High Entropy in one channel. This points to a trend towards higher activation for non-repetition sequences, and also higher activation for High Entropy. However, we did not find an overall difference between the two grammars across channels. Neither did we find an overall difference between the low and high entropy conditions. These results suggest that 6-month-olds are able to process both the repetition and the non-repetition patterns, and the processing costs are the same for both patterns. Our findings are the first to reveal a developmental change in language acquisition

[12] This chapter is a modified version of a manuscript in preparation: Radulescu, S., Wijnen, F., Avrutin, S., and Gervain, J. (2021) Same Processing Costs for Repetition and Non-Repetition Grammars in 6-month-olds: An fNIRS Study

between the age of 6 months and birth, when discrimination between repetition and non-repetition patterns was found (Gervain et al., 2008). This ability to encode a sequence of different syllables, may support 6-month-olds' growing language skills, e.g. the beginning of word learning. Thus, our study contributes to a better understanding of the developmental trajectory and the nature of the sameness/difference representations, which underlie the building blocks of rule learning in language.

1. Introduction

Infants can learn regularities from their linguistic input. In order to account for the detection and encoding of simple repetition-based grammars (e.g. ABB "bu-ra-ra"), previous studies with infants and adults proposed two qualitatively different mechanisms: abstract rule learning, based on symbolic encoding of variables (Marcus, Vijayan, Rao, & Vishton, 1999), and a low-level perceptual mechanism, based on automatic sensitivity to repetitions (Endress, Nespor, and Mehler, 2009). Exploring the development of these mechanisms is necessary for a better understanding of infants' ability to represent repetition-based relations or sameness (ABB) and non-repetition relations or difference (ABC), given that these types of representations contribute to the building blocks of structure acquisition.

From a series of artificial grammar studies, Marcus et al. (1999) concluded that 7-month-olds recognize and generalize repetition-based structures, like AAB strings such as "*le-le-di*", "*ko-ko-ba*", based on the findings that infants were able to discriminate new strings (which had not been presented during familiarization) with the same AAB structure from novel strings with a different structure (ABB or ABA). This shows that infants could not simply rely on rote memorization or transitional probabilities between specific familiar items. Marcus et al. argue that the underlying mechanism supporting rule learning is an innate abstract symbol-manipulating mechanism that operates on variables. Whether the ability to encode patterns is supported indeed by an innate abstract mechanism remains largely underspecified and, thus, hotly debated (Aslin & Newport, 2012; 2014; Frost & Monaghan, 2016; Radulescu, Wijnen, & Avrutin, 2019). Nonetheless, these results point to 7-month-olds' ability to represent abstract repetition-based relations, since they were able to generalize the rule to novel instances.

Gerken (2006) took a step further towards understanding the factors that trigger rule learning in infants, by showing that the nature of the representation (generalization) that learners form depends crucially on the statistical properties displayed by the input. By modifying the design used by Marcus et al. (1999), Gerken (2006) asked whether 9-month-olds presented with two different subsets of the strings used by Marcus et al. (1999) would generalize repetition-based AAB regularities. Thus, she had one group of infants exposed to four AAB strings ending in different syllables (*je/li/di/we*) and another group to four AAB strings ending only in *di*. The second group only made a narrow generalization to strings that ended in "*di*", i.e. AA*di,* while the first group made a broader generalization to AAB. Gerken (2006) interpreted these results to point to an effect of stimulus variability on the type of generalization that learners make: learners in the second group had both equally plausible generalizations at hand, i.e. the narrow "ends in *di*" and the broad AAB rule, but, since they did not have any clear evidence that strings can end in any other syllable, they inferred the maximally reliable rule "ends in *di*". In a later study, Gerken (2010) exposed 9-month-olds to the same "ends in *di*" as in Gerken (2006), but at the end of the exposure phase three strings ending in "*je/we/li*" were added. The learners made the broader (AAB) generalization in this case. The author interpreted these results to show that it is not the mere number of items in the input, but the logical structure of the input that drives a "rational" decision-making process, which resembles a Bayesian type of learning based on incrementally updating hypotheses as supported by the direct evidence provided by the input.

Radulescu et al. (2019) challenge this interpretation and argue that, formally, none of these groups of infants saw direct evidence that strings could end in a new syllable, except for "*je/li/di/we*" in the higher variability group, and "*di*" in the lower one. Nonetheless, learners exposed to the higher variability input made a broader *category-based generalization* (AAB), instead of sticking to the narrower *item-bound generalization* (AA*di*). Radulescu et al. (2019) argue that the evidence proposed by Gerken (2010) in favor of "rational" learners who generalized based on the logical structure of the input is not sufficient or decisive: if presented with the three strings ending in "*je/li/di/we*" at the beginning of the 2-minute-familiarization, Radulescu et al. (2019) suggest that the learners might "forget" them, and only update their model based on the more strongly evidenced and recent "ends in *di*" input.

While these findings among others (Gómez, 2002) point to an effect of input variability on rule induction in infants, the necessary amount and nature of input variability remains unclear: some studies show that at least three or four examples are needed for infants to generalize (Gerken, 2006; 2010; Gerken & Bollt, 2008; Peterson, 2011), but Gerken et al. (2015) claim that a single example suffices for generalization. However, it seems that it is not mere variability that is critical, but a specific pattern of variable input (Gómez, 2002).

We proposed *input entropy* to quantify this specific pattern of variable input, and we put forth a new information-theoretic entropy model to rule induction (Radulescu et al., 2019), which employs the Shannon's noisy-channel coding theory (Shannon, 1948). While in Radulescu et al. (2019), we tested the model on repetition-based grammar learning in adults, the aim of the present study is to extend the model to rule induction in infants. In Radulescu et al. (2019), we distinguished between two qualitatively different types of rule induction (generalizations): *item-bound generalization* and *category-based generalization,* by following suggestions from previous conceptualizations in the literature (Gómez & Gerken, 2000). While *item-bound generalizations* are defined as generalizations bound to specific items present in the input (e.g. "every string ends in *di*" generalization made by the low variability group in Gerken, 2006), *category-based generalizations* are operations beyond specific items in the input, spanning over novel instances (e.g. the AAB generalization made by the high variability group in the same study).

In short, and simplifying for now, the main hypothesis of our entropy model is that rule induction is an encoding mechanism driven by an external factor – the statistical properties of the input, i.e. *input entropy,* which interacts with an internal factor – the brain's ability to encode the input under conditions of finite encoding capacity (i.e. *channel capacity*). The encoding capacity is defined as *channel capacity,* in information-theoretic terms, that is the finite rate of information encoding (entropy per unit of time), which might be supported by various cognitive capacities, e.g. memory capacity, in psychological terms. Our entropy model hypothesizes that *item-bound generalization* and *category-based generalization* are not independent mechanisms. Rather, they are outcomes of one phased mechanism that *gradually* moves from memorized combinations of items to a high-specificity encoding (*item-bound generalization*), and eventually to a high-generality encoding (*category-based generalization*). Specifically, if *input entropy* is lower than the available *channel capacity,* the input can be encoded using high-specificity *item-bound generalization*, while an increase in *input entropy* gradually shapes *item-bound generalization* into *category-based generalization*, in order to avoid exceeding the *channel capacity*. It follows that a reduced *channel capacity*, which is assumed to be supported by cognitive capacities that are not yet fully matured in the developing brain, might support the transition to *category-based generalization* under conditions of relatively low input entropy, i.e. lower than the input entropy that adults might need.

Indeed, learners' cognitive capacities were previously proposed as the internal factor that drives rule induction: the classical *Less-is-More* hypothesis (Newport, 1990; 2016) and subsequent related studies (Hudson Kam & Chang, 2009; Hudson Kam & Newport, 2005; 2009) propose, and show some evidence for children's higher tendency to move away from the statistical specificity of the input and generalize rules from the input as compared to adults. Adults were shown to learn and rather stick to the probability distributions displayed by the input (i.e. probability matching) instead of generalizing. However, when exposed to a memory overloading and noisier input, i.e. high variability (Hudson Kam & Newport, 2009; Hudson Kam & Chang, 2009), adults were also shown to generalize. Thus, under the *Less-is-More* hypothesis, children's tendency to generalize is assumed to be driven by their incomplete cognitive development (maturational constraints – Newport 1990, 2016), more specifically by memory constraints (children's overall lower memory capacity – Cowan, 1997; Gathercole, 1998).

However, from a developmental perspective, the question of when in infancy such generalizing abilities and, specifically, the ability to represent relations of repetition (sameness) and non-repetition (difference) develop is still an open question. While there is evidence for the ability to learn repetition-based grammars from birth (Gervain, Berent, & Werker, 2012; Gervain, Macagno, Cogoi, Peña, & Mehler, 2008) and throughout the first year of life (Gerken, 2006; Marcus et al., 1999), the developmental trajectory and the processing demands of forming representations of repetition/non-repetition (sameness/difference) have yet to be thoroughly investigated. Brain maturation, which supports the development of memory and other cognitive capacities, is thought to play a crucial role in grammar learning mechanisms, i.e. maturational constraints on rule learning (Newport, 1990; 2016). In the current study, we extend to infants' rule induction our entropy model proposed for adults' rule induction in Radulescu et al. (2019), which puts together both external factors (*input entropy*) and internal factors (*channel capacity*) in one consistent account.

2. An entropy model for rule induction in infants

2.1 A brief introduction to our entropy model and previous findings

Radulescu et al. (2019) propose a new information-theoretic entropy model for rule induction, which offers a more refined formal approach to the *Less-is-More hypothesis* (Newport, 1990; 2016). The basic hypothesis of this model is that the factors triggering the transition from *item-bound* to *category-based generalization* are *input entropy*, and our brain's finite encoding rate, i.e. *channel capacity*. We use the concepts and formulas for *entropy* and *channel capacity* as they were introduced and mathematically demonstrated by Shannon (1948).

Entropy is as a function of the number of items and their probability distribution. For a random variable *X*, with *n* values {x_1, x_2 … 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 $p(x_i)$ is the probability that x_i occurs[13]. Entropy is used here to capture and describe a property of the input – *a specific pattern of variability*, and as a measure of this property, i.e. a measure of *input variability.* Entropy (measured in bits) captures the unique dynamics between two factors (number and probability distribution of items) that defines a specific pattern of variability proposed by Radulescu et al. (2019) to be relevant for rule learning.

The other factor used in the entropy model is another information-theoretic concept, i.e. c*hannel capacity*, which is defined by the amount of entropy that can be transmitted through the channel per unit of time (Shannon, 1948). *Channel capacity* is used to model the finite encoding rate of the information encoding system, i.e. the amount of entropy that can be encoded per unit of time (bits/s). Previous memory capacity studies showed that our capacity to encode specific items and relations between them is finite (Baddeley, Eysenck, and Anderson, 2015; Cowan, 2005). Thus, the dynamics between *input entropy* and the finite *channel capacity* are hypothesized to drive different forms of information encoding employed to encode the complexity of a given input, as follows.

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

According to our entropy model, *item-bound generalization* and *category-based generalization* are outcomes of the same information encoding mechanism that *gradually* goes from a high-specificity form of encoding (*item-bound generalization)* to a high-generality encoding (*category-based generalization)*, as triggered by the interaction between *input entropy* and the finite encoding capacity of the brain.

If the input entropy is low – that is below or matches the *channel capacity*, then the input can be encoded using an encoding method that matches the input statistical structure, i.e. the probability distribution of the specific items in the input. Thus, the items with their specificity can be encoded and transmitted through the channel at the available channel rate (i.e. amount of entropy per unit of time), and stored by *item-bound generalization,* that is probability matching to the input. If the *input entropy* is higher than the finite *channel capacity* of the encoding system, the encoding rate cannot exceed the channel capacity (Shannon, 1948). As a consequence, this essential design feature of the *channel capacity* "forces" the information processing system to re-structure the information to *gradually – bit by bit* – shape the *item-bound generalization* into *category-based generalization*.

As we argued in Radulescu et al. (2019), re-structuring the information entails (unconscious) re-observing the item-specific features and structural properties of the input and identifying specific features that are the same or different across items. As a result, the information can be compressed by *gradually* reducing the number of specific different features that individual items are coded for (i.e. by erasing or "forgetting" insignificant differences between the items, that is low probability features). As a result of reducing ("forgetting") the specific features, items are grouped in "buckets" (i.e. categories) based on the same non-specific shared features (Radulescu et al., 2019). This would model the step-by-step birth of abstract categories: such as *AAB* or *ABB* patterns, which allow for novel items to be included in these categories, based on relations of sameness and difference.

In developmental terms, an increase in *channel capacity*, (e.g. resulting from growth/development of cognitive capacities), that is an increase in the amount of entropy that can be processed and encoded per unit of time, reduces the need for, and thus the tendency to move to a higher-generality *category-based* form of encoding. Thus, infants are hypothesized to have a higher tendency than adults to move towards *category-based generalization* from exposure to a lower *input entropy*.

As we argued in Radulescu et al. (2019), sensitivity to entropy entails a sensitivity to a specific pattern of variability in the input, which is given by the degree of sameness/difference between items and their features, and also their probability distribution, which assigns them significance. The more differences are encoded between specific items (that is many different specific features encoded for each item – measured in bits of information), the higher the degree of specificity of the encoding (i.e. *item-bound* specificity). Conversely, since the *channel capacity* places an upper bound on the number of bits encoded per unit of time, a reduction – *"gradual forgetting"* – of the encoded differences highlights a higher degree of sameness, hence the lower the degree of specificity and the higher the degree of generality. Entropy captures these dynamics of sameness vs difference, and quantifies it in bits of information.

This model was first proposed and tested on rule induction in adults in Radulescu et al. (2019), where in two artificial grammar experiments we exposed adults to six versions of a repetition-based XXY grammar with different entropy levels. Results showed that adults' tendency to move from *item-bound* to *category-based generalization* increased *gradually* as a function of increasing *input entropy*, as predicted by our entropy model. In the current study, we extend the entropy model to rule induction in infants, based on the rationale that infants' reduced *channel capacity* is the hypothesized driving factor for a high tendency to move towards *category-based generalization* when exposed to relatively low input entropy, as compared to adults.

2.2 Detecting repetition (sameness) and non-repetition (difference) in infancy

Do infants also move gradually from *item-bound* to *category-based generalization* as a factor of increasing *input entropy,* as previous findings seem to indicate (Gerken, 2006)? Since *channel capacity* in infants is hypothesized to be lower than in adults, given that infants' cognitive capacities are yet to be fully developed, the empirical investigation of this question is deeply relevant to extending our entropy model to infants. A comprehensive answer to this question requires firstly a conceptual argumentation of the type of evidence that can be considered to show a gradual transition from *item-bound* to *category-based generalization* in infants, which is currently missing in the literature. In the present study we start by asking two closely related and more specific questions. How do infants process repetition-based (ABB) regularities as compared to non-repetition-based (ABC) regularities? Also, does *input entropy* have a positive effect on the processing costs? Before presenting the design and the rationale of our study, in this section we discuss previous relevant studies that addressed similar questions.

In an optical imaging study, Gervain et al. (2008) probed whether neonates can discriminate repetition-based grammars (ABB: "mu-ba-ba" and ABA: "ba-mu-ba") from random controls (ABC: "mu-ba-ge"). The study employed functional NIRS to measure the hemodynamic response as a proxy for the neural activity associated with processing ABB sequences vs ABC sequences. The results showed significantly greater increase in oxyHB for ABB than for ABC in both left and right temporal areas. However, in the second experiment, which tested the response to ABA vs ABC sequences with the same design and procedure, there was no difference between the amplitudes of the response to ABA vs ABC sequences, although both sequences gave rise to a response that was significantly greater than baseline. The authors interpreted these results to show that neonates possess an innate ability to detect repetition-based sequences (ABB), however not in non-adjacent position (ABA).

In three fNIRS experiments, Gervain et al. (2012) investigated the ability of the newborn brain to detect identity relations both in a sequence-initial repetition (AAB) grammar and a sequence-final repetition grammar (ABB). In the first experiment, they found greater oxyHB and smaller deoxyHB responses for AAB sequences, compared to ABC sequences, which showed that newborns were able to discriminate repetition-based (AAB) sequences from random controls (ABC). The second experiment compared newborns' response to ABB vs AAB sequences in an alternating/non-alternating block design (ABB sequences alternating with AAB sequences in half of the blocks, while the other half of the blocks presented either AAB or ABB sequences). They found a larger response to the non-alternating blocks in the left frontal areas, indicating that newborns could discriminate the two grammars. In the third experiment, the authors compared newborns' response to ABB vs AAB sequences in a simple block design, resembling the design used by Gervain et al. (2008), and found no difference between the responses (i.e. they found similar canonical hemodynamic responses in the bilateral temporal and left frontal areas). Since NIRS measures the hemodynamic response as a signature of metabolic effort related to neural processing, the authors interpreted the similar responses as possible evidence for similar processing costs associated with both types of sequences (ABB and AAB), with no advantage or preference for any of the structures.

In a behavioral experiment that resembled the design by Marcus et al. (1999), Gervain and Werker (2012) looked at the ability of 7-month-olds to discriminate adjacent-repetition (ABB) and non-adjacent repetition (ABA) sequences from non-repetition ABC controls. One group was familiarized with the ABB grammar, and another group with the ABA grammar, and in the test phase they were presented with novel sequences of their familiarization grammar and ABC sequences. They found significantly longer looking times to the ABC controls than to their respective familiarization sequences, showing that 7-month-olds have the ability to learn and represent both adjacent (ABB) and non-adjacent (ABA) repetition-based grammars. In a follow-up experiment, Gervain and Werker (2012) found that when given no previous familiarization with any of the grammars, 7-month-olds do not show any spontaneous preference for the repetition-based or non-repetition controls, as it would be predicted by an automatic perceptual repetition-detector theory (Endress et al., 2009) . This study ads an important piece of the puzzle to the study by Marcus et al. (1999), by bringing evidence that 7-month-olds do not only encode representations of sameness to help them encode position-dependent *same-same* relations, i.e. AAB vs ABB or ABA, but they can also perceive such relations as different from non-repetition sequences (ABC), that is from relations of difference between items.

Wagner, Fox, Tager-Flusberg, and Nelson (2011) addressed the question of repetition-based grammar learning from a developmental point of view in an fNIRS study that compared 7-month-olds to 9-month-olds in their ability to discriminate ABB and ABC sequences, but only found a marginally significant interaction between condition (ABB, ABC) and age (7-month-olds, 9-

month-olds): there was a larger negative deoxyHB response for ABB sequences in the younger group, while the 9-month-olds showed the opposite, i.e. a larger negative deoxyHB response for ABC sequences. The authors interpreted these findings as a possible shift in development with regard to the brain's response to repetition and non-repetition grammars during the first year of life. However, the exact nature, the processing demands and the stage in development when a shift emerges remain underspecified.

Summarizing, some studies (Gervain et al., 2008; Gervain et al., 2012) show that the neonate brain possesses an innate sensitivity to repetition-based relations, evidenced by higher activation (i.e. higher amplitude of the hemodynamic response) to repetition grammars (ABB and AAB) compared to non-repetition grammars (ABC), as well as compared to baseline. By contrast, no cortical area in the newborn brain responds to sequences of three different syllables (ABC), when compared to baseline. Such evidence could point to an innate ability to perceive relations of repetition (sameness), while the ability to detect and process relations of difference (ABC) may not yet be developed. Taking this evidence and interpretation into consideration together with the above-mentioned experiments by Gervain and Werker (2012), there seems to be a developmental change between birth and the age of seven months. Specifically, newborns are born with an ability to perceive and represent repetition-based grammars (sameness), while 7-month-olds show (at least behaviorally, by longer looking times) an interest for relations of difference as well. However, since this does not constitute evidence that 7-month-olds can encode relations of difference, further research is needed in order to specify how young infants process repetition-based grammars as compared to non-repetition grammars.

Another line of research with infants showed a positive effect of input variability on rule learning (Gerken, 2006; Gómez, 2002). However, *how* exactly stimulus variability plays a role in the ability to form representations of sameness/difference has yet to be clearly specified. Nonetheless, these studies and others (Gerken & Bollt, 2008; Gómez, 2002) clearly show that there is a *gradient of generalization* depending on the statistical properties of the input (Aslin and Newport, 2012; 2014). In addition, they also show that little input variation is needed for infants to move from narrow *item-bound generalization* to broader *category-based generalization*. In fact, Gerken, Dawson, Chatila, and Tenenbaum (2015) suggest that variability is not needed for infants to make a broader generalization (AAB). Specifically, they found that 9-month-olds only need one example to generalize, if the input presents them with a repetition pattern that is "surprising" ("*le-le*"), since their prior language model – English, in this case – does not have such repetition patterns as a common feature.

In an fNIRS study that followed up on Gervain et al. (2008), Bouchon, Nazzi, & Gervain (2015) investigated the role of stimulus variability in the newborns' ability to discriminate repetition-based ABB sequences from random ABC controls. In this study they used less unique trisyllabic sequences – 24 different sequences – as compared to Gervain et al. (2008), which presented newborns with 280 different sequences, so that no sequence was repeated throughout the entire experiment. Unlike Gervain et al. (2008) who found a repetition enhancement effect for the ABB sequences, Bouchon et al. (2015) found a repetition enhancement in response to the ABC controls. More specifically, they found different time-dependent dynamics of the newborns' responses to the two different grammars. While the response to the repetition-based ABB remained constant over time after a small initial increase, the amplitude of the response to ABC (in the left fronto-temporal cortex) increased over the time of the experiment. The authors interpreted the increased amplitude of the response to ABC as evidence of the effect of stimulus variability, i.e. the effect of less redundancy in the ABC controls (vs the more repetitive ABB stimuli). The authors argue that less redundancy might have encouraged learning based on memorization of the material, thus eliciting higher processing costs indicated by an increased neural effort for the non-repetitive ABC controls.

3. Design and rationale of the present study with infants

We set out to find the effect of *input entropy* on infant rule learning, in order to probe whether and how *input entropy* impacts rule induction in infants, i.e. if infants' tendency towards *category-based generalization* increases as a function of increasing *input entropy*. As discussed above, previous research into infants' generalization mechanisms showed that variability plays a role in infant rule learning (Bouchon et al. 2015; Gerken, 2006; Gómez, 2002). However, input variability has not been systematically quantified by using entropy (or other information-theoretic measures) in previous studies with infants, while the questions of *how* and *why* input variability should have an effect are still largely unanswered. The goal of the present study is to fill this gap and to offer a consistent information-theoretic account to *how* and *why* developmental changes in cognitive capacities should be a driving factor for rule induction, previously formulated as maturational constraints (Newport, 1990; 2016).

This study looks at whether young infants process repetition grammars and non-repetition grammars in a similar way to newborns, or whether they follow a different developmental pattern. One insufficiently explored issue is how young infants process repetition-based grammars as compared to non-repetition grammars. Newborns seem to process adjacent-repetition grammars (ABB/AAB) differently from a non-repetition ABC grammar (Gervain et al., 2008; Gervain et al., 2012) evidenced by a significantly higher activation for the ABB/AAB strings than for the random controls (ABC). However, *whether* and *how* repetition vs non-repetition relations are processed by young infants is still largely underspecified. Specifically, we ask whether the processing costs of repetition grammars differ from the processing costs of non-repetition grammars?

In order to address our research questions, we tested whether and how 6-month-old infants process repetition-based linguistic regularities (ABB, e.g. "bu ra ra") as compared to non-repetition sequences (ABC, e.g. "bu fa zo") manipulating the entropy (low vs high) of the stimuli using near-infrared spectroscopy (NIRS). We used trisyllabic sequences, resembling the stimulus material used by Gervain et al. (2008) and Bouchon et al. (2015), but with three major differences in 1) stimulus entropy, 2) the age of infants and 3) the experimental design.

Firstly, in order to further the investigation of the effect of input variability started by Bouchon et al. (2015), we created two different input entropy conditions – low entropy and high entropy – with the low entropy level below the low complexity, and the high one above the entropy level that we calculated for the stimulus material used in Bouchon et al. (2015). However, we chose both entropy levels to be below the extremely high entropy we calculated for the stimulus material used in Gervain et al. (2008).

Secondly, we decided to test 6-month-old infants rather than newborns based on the following rationale. At this age, infants learn the most basic grammatical properties of their native language (Gervain et al., 2012). In addition, infants start learning their first words and develop a word-learning capacity that goes beyond specific associations to learning words that refer to categories of objects (Tincoff & Jusczyk, 2012; Bergelson & Swingley, 2012). Thus, this is a stage where it becomes crucial for language development to be able to encode both relations of repetition (sameness) and non-repetition (difference), which are essential for the ability to generalize.

Thirdly, in order to ensure that the experiment was suitable for 6-month-olds in terms of duration (to keep their attention for the entire duration of the experiment), we used a similar interleaved block design as Gervain et al. (2008), but we had significantly reduced number of blocks per condition (i.e. 3 blocks vs 14 blocks), with a total of six blocks (3 ABB and 3 ABC) per entropy condition. Thus, the total testing time (8 minutes) was much shorter than the one used with sleeping newborns (20 – 25 min) in Gervain et al. (2008) and the one (14 – 15 min) in Bouchon et al. (2015).

We predicted that we would find higher activation (i.e. higher amplitude of the hemodynamic response) compared to baseline for both ABB and ABC sequences. This would show that 6-month-olds are able to process both repetition and non-repetition grammars. In addition, we also predicted a different level of activation for the ABB sequences compared to the ABC sequences, which could constitute evidence for infants' ability to discriminate between repetition and non-repetition sequences as a possible indication of grammar learning. Thirdly, we predicted that this difference in activation levels for the ABB vs ABC sequences would be larger in the high entropy condition than in the lower entropy condition. Such a difference would show an effect of input entropy on processing, and possibly, learning the grammars.

4. Materials and Methods

4.1 Participants

Twenty-one full-term, French-exposed, healthy 6-7-month-old infants (M age: 6.55 months, age range: 6–7 months; 10 boys, 11 girls) were included in the analyses. An additional 17 infants were tested, but were excluded from the analysis due to fussiness, failure to start/complete the procedure, including nonstarter babies who refused the cap to be placed on their head, or due to insufficient analyzable data (see below). All parents gave informed written consent before the beginning of the experiment. The study was approved by the CERES ethics committee of the Université Paris Descartes (Université de Paris as of January 2020) under number 2011-13.

4.2 Stimuli

Infants were exposed to a repetition-based ABB grammar (e.g. "bu ra ra") and random controls (ABC, e.g. "bu fa zo"), similar to the stimuli used by Gervain et al. (2008) and Bouchon et al. (2015). Both grammars generated non-sense trisyllabic sequences of CV syllables and were matched for their phoneme repertoire (i.e. 6 consonants and 6 vowels – Table 1), prosody and transitional probabilities between adjacent syllables (0.33). Sequences were synthesized using the MBROLA diphone database with the French fr4 female voice (Dutoit, Pagel, Pierret, Bataille, and Vreken, 1996), in a monotonous pitch (200Hz) and the same duration of phonemes (i.e. consonants: 120ms, vowels: 150ms).

Using these consonants and vowels, we created two different sets of stimuli with different input entropy: a low entropy condition (3.17bits) and a high entropy condition (4.17bits). The overall cross-condition entropy was 3.67bits. The low entropy grammar contained 9 unique CV syllables (Table 1), which were concatenated in 9 ABB sequences and 9 ABC sequences, each of them occurring twice within the entire duration of the experiment. The high entropy grammar contained 18 CV syllables (Table 1), which were concatenated in 18 ABB sequences and 18 ABC sequences, each occurring once within the duration of the experiment.

6 Cs	6 Vs	Low Entropy: 9 syllables	High Entropy: 18 syllables	
b	i	bi	bi	Zi
f	e	fa	fa	pe
R	y	Ry	Ry	go
g	a	ge	ge	Re
Z	u	Zo	Zo	Za
p	o	pu	pu	fi
		bu	bu	by
		fy	fy	gu
		Ra	Ra	po

Table 1. The phonemes and syllables used in Low Entropy and High Entropy conditions

For the entropy calculations, we used the same entropy calculation model as in Radulescu et al. (2019), which is a more refined method based on a method proposed by Pothos (2010) for finite-state grammars (see Table 2 for complete calculations).

Low Entropy	High Entropy

H[beginA]=H[9]= -Σ[0.111*log0.111] = 3.169
H[AB] = H[9]= 3.169
H[BB] = H[9] = 3.169
H[Bend] = H[9] = 3.169
H[beginAB] = H[9] = 3.169
H[ABB] = H[BBend]= H[9] = 3.169
H[bigram] = 3.169
H[trigram] = 3.169
H[total] = $\frac{\mathbf{H[bigram]+H[trigram]}}{\mathbf{2}}$ **= 3.169**
	H[beginA]=H[18]=
	-Σ[0.055*log0.055] = 4.169
	H[AB] = H[18]= 4.169
	H[BB] = H[18] = 4.169
	H[Bend] = H[18] = 4.169
	H[beginAB] = H[18] = 4.169
	H[ABB] = H[BBend]= H[18] = 4.169
	H[bigram] = 4.169
	H[trigram] = 4.169
	H[total] = $\frac{\mathbf{H[bigram]+H[trigram]}}{\mathbf{2}}$ **= 4.169**

Table 2. Entropy values

4.3 Procedure

All participants were presented with both ABB and ABC blocks in an interleaved design, and to both low entropy and high entropy conditions sequentially in a counter-balanced order (Figure 1A). Each block contained 6 ABB or ABC sequences separated by a brief pause of either 0.5s or 1.5s (duration was chosen randomly). Blocks were separated by a 20s or 25s pause chosen randomly to prevent phase-locked brain responses. Block order and condition order were pseudo-randomized and counter-balanced across infants.

Each entropy condition contained 3 ABB blocks and 3 ABC blocks, with a total duration of 3.6 minutes per condition. Between the two entropy conditions, infants listened to a music track of 37s, in order to provide them with a short pause and an auditory stimulation of an entirely different nature between the two entropy conditions. The total testing time for each infant was 8 minutes (Figure 1A).

Infants were tested with a NIRx NIRScout 8-16 machine in a quiet dimly lit testing booth at the BabyLab of the Université Paris Descartes. The optical sensors were inserted into a stretchy cap and placed on the infants' head using surface landmarks (nasion, and the preauricular points), targeting the language areas in the bilateral temporal, frontal and parietal cortices (Figure 1B). These areas match those that responded to reduplication in speech in newborns (Gervain et al. 2008). We approximated the cortical regions underlying our NIRS channels following Lloyd-Fox et al. (2014) and Abboub, Nazzi, and Gervain (2016), using age-appropriate structural MRIs and stereotaxic atlases (Fillmore, Richards, Phillips-Meek, Cryer, & Stevens, 2015; Kabdebon et al., 2014). The position of optodes was measured with respect to the nasion and tragi for each participant and, together with photographs of the optode positions, were used to localize the optodes on a structural whole head MRI image. The locations were then projected down onto the cortical surface to identify the regions underlying the NIRS channels for each infant. A channel was then labeled according to the localization found in the majority of participants. Accordingly, channels 1, 2, 4, 5 and 13-16 query the frontal lobe, channels 3, 6, 17 and 19 are positioned over the temporal lobe, channels 7, 10, 12, 18, 20 and 23 are parietal, whereas channels 9 and 21 span the temporal and parietal lobes.

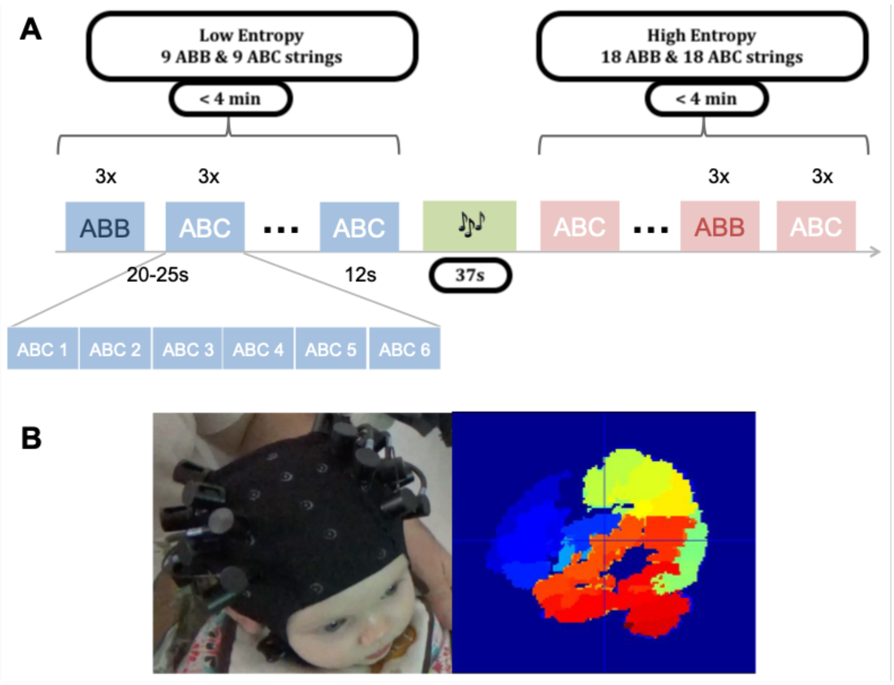


Figure 1. Block design (A) and probe placement (B)

During testing, infants were seated on a caregiver's lap. The stimuli were presented aurally using E-Prime through speakers placed on the right and left side of a computer screen located in front of the infants at approximately 80cm. During stimulus auditory presentation, a cartoon was playing on a screen. Caregivers were instructed not to talk to their infant or orient their behavior.

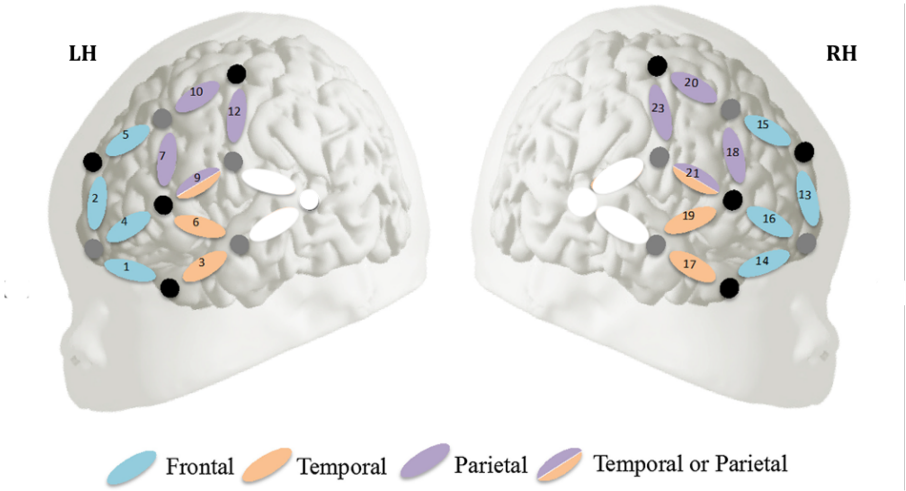


Figure 2. Probe placement with channel labeling. Grey circles represent sources and black circles represent detectors of the signal.

4.4 Data analysis

The NIRS machine measured the intensity of the transmitted light, from which concentration changes of oxygenated hemoglobin (oxyHb) and deoxygenated hemoglobin (deoxyHb) were calculated using the modified Beer-Lambert Law. To eliminate noise (e.g., heartbeat) and overall trends, the data were band pass-filtered between 0.01-0.7Hz. Movement artifacts, defined as concentration changes above 0.1 mmol*mm over two samples, were removed by rejecting block-channel pairs in which artifacts occurred. For valid, non-rejected blocks, a baseline was linearly fitted between the means of the 5s preceding the onset of the block and the 5s starting 15s after offset of the block (in accord with general practice in NIRS studies – Lloyd-Fox, Blasi, & Elwell, 2010). Infants were videotaped during the experiment.

Statistical analyses were carried out on the average of both oxyHb and deoxyHb concentrations recorded in a time window starting at 9 s after stimulus onset and until 22 s after stimulus onset, i.e. containing the 10-second-relaxation period.

5. Results

The grand average results are presented in Figure 3. The figure shows the oxyHb and deoxyHb concentration changes averaged across all blocks of each condition and across all infants. Figure 4 shows the results for the ABB vs ABC comparison across all blocks collapsed over entropy conditions, while Figure 5 shows the results for the Low vs High Entropy comparison across all blocks collapsed over grammar conditions.

Firstly, we averaged the oxyHb and deoxyHb concentrations across all blocks in each channel, for each grammar in each entropy condition. We performed channel-by-channel t-tests for the mean oxyHb and deoxyHb changes for the following comparisons: ABB vs baseline, ABC vs baseline, ABB vs ABC across Entropy conditions and in each Entropy Condition, ABB in Low Entropy vs High Entropy, ABC in Low Entropy vs High Entropy.

Results revealed significant activation for ABB compared to baseline in channel 6 (ABB, oxyHb, *p* = .049), channel 17 (ABB, deoxyHb, *p* = .026), and significant activation for ABC compared to baseline in channel 2 (ABC, oxyHb, *p* = .020), channel 17 (ABC, deoxyHb, *p* = .018).

Further, channel-by-channel t-tests yielded a significant difference for the following ABB vs ABC comparisons collapsed over Entropy conditions: channel 2 (ABB < ABC, oxyHb, *p* = .034), channel 20 (ABB > ABC, oxyHb, *p* = .048). Also, results yielded a significant difference for the following comparisons with Entropy conditions: channel 5 (ABC in Low Entropy < ABC in High Entropy, deoxyHb, *p* = .039), channel 10 (ABC in Low Entropy < ABC in High Entropy, deoxyHb, *p* = .028), channel 12 (ABC in Low Entropy < ABC in High Entropy, deoxyHb, *p* = .044); channel 3 (ABB in Low Entropy < ABB in High Entropy, oxyHb, *p* = .041), channel 4 (ABB in Low Entropy > ABB in High Entropy, oxyHb, *p* = .033); channel 2 (ABB < ABC in High Entropy, oxyHb, *p* = .012).

Given that overall oxyHb yielded more significant activation than deoxyHb, in line with previous findings (Gervain et al., 2008), and also given that oxyHb is most commonly and robustly employed in the literature for infants (Aslin, Shukla, & Emberson, 2015), we ran further analyses on oxyHb as a better predictor.

As the order of the Entropy conditions did not yield a significant main effect in preliminary analyses, we collapsed over the order in the analyses. Next, we grouped the 20 channels according to hemisphere (left hemisphere – LH, right hemisphere – RH, Fig. 2) and ROI (Frontal, Temporal, Parietal, Fig. 2), and averaged the oxyHb across all blocks for each channel. Firstly, a repeated measures ANOVA with the main within-subjects factors Grammar (ABB/ABC), Entropy (Low/High), and Entropy * Grammar interaction, Entropy * Grammar * Hemisphere (Left/Right) interaction, and Grammar * Hemisphere interaction was run on average oxyHb concentration across channels to evaluate whether the two grammars are processed differently in the two entropy conditions and in the two hemispheres. No significant effects or interactions were found. Secondly, another repeated measures ANOVA with the main within-subjects factors Grammar (ABB/ABC), Entropy (Low/High), and Entropy * Grammar interaction, Entropy * Grammar * ROI (Frontal, Temporal, Parietal) interaction, and Grammar * ROI interaction was run on average oxyHb concentration across channels to evaluate whether the two grammars are processed differently in the two entropy conditions and in the three ROIs. No significant effects or interactions were found.

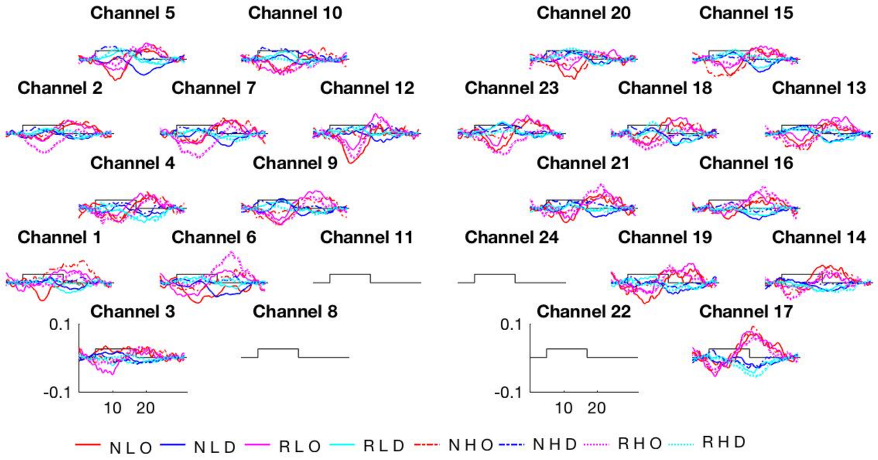


Figure 3. Grand average results. The concentration changes of oxy- and deoxyHb were averaged across all blocks for each condition and for each channel. The x-axis shows time (seconds), the y-axis shows hemoglobin concentration (mmol x mm). The rectangle along the x-axis indicates time of stimulation. The continuous red and blue lines in the graphs represent oxyHb (O) and deoxyHb (D) concentrations, respectively, in response to the ABC (N=non-repetition) grammar in the Low Entropy (L) condition. The continuous magenta and cyan lines represent oxyHb and deoxyHb concentrations, respectively, in response to the ABB (R=repetition) grammar in the Low Entropy condition. The dashed red and blue lines represent oxyHb and deoxyHb concentrations, respectively, in response to the ABC (N=non-repetition) grammar in the High Entropy (H) condition. The dashed magenta and cyan lines represent oxyHb and deoxyHb concentrations, respectively, in response to the ABB (R=repetition) grammar in the High Entropy condition. The time line on the x-axis shows the following sequence of events: 5 s time-window before the onset of the block, block presentation (12 s), and the between-block silences (20 s or 25 s), giving a total duration of 37–42 s.

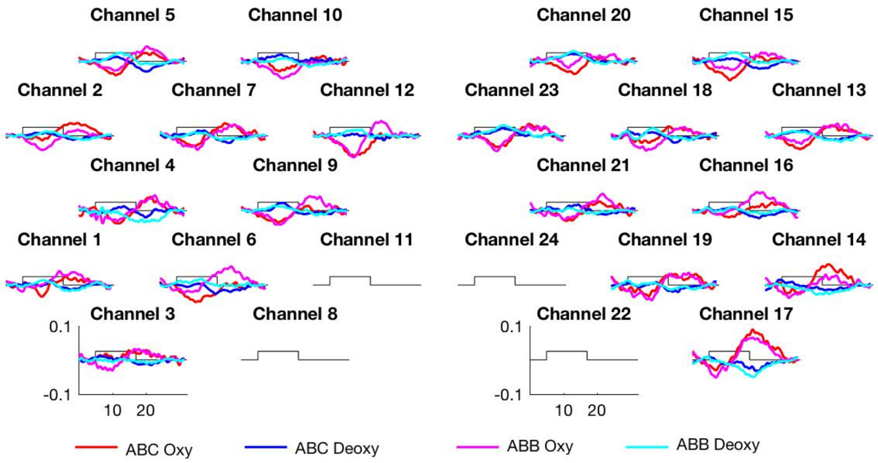


Figure 4. Grand average results for ABB vs ABC grammars. The concentration changes of oxy- and deoxyHb were averaged across all blocks for each condition and for each channel. The x-axis shows time (seconds), the y-axis shows hemoglobin concentration (mmol x mm). The rectangle along the x-axis indicates time of stimulation. The continuous red and blue lines in the graphs represent oxyHb (O) and deoxyHb (D) concentrations, respectively, in response to the ABC grammar. The continuous magenta and cyan lines represent oxyHb and deoxyHb concentrations, respectively, in response to the ABB grammar.

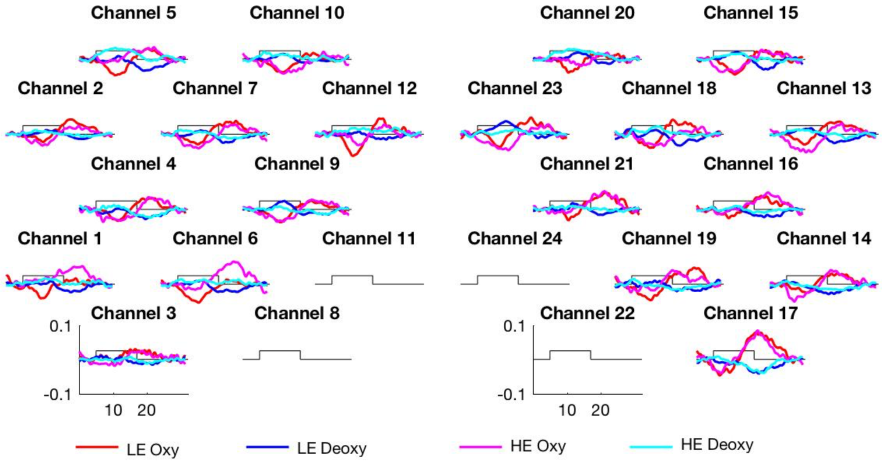


Figure 5. Grand average results for Low vs High Entropy. The concentration changes of oxy- and deoxyHb were averaged across all blocks for each condition and for each channel. The x-axis shows time (seconds), the y-axis shows hemoglobin concentration (mmol x mm). The rectangle along the x-axis indicates time of stimulation. The continuous red and blue lines in the graphs represent oxyHb (O) and deoxyHb (D) concentrations, respectively, in response to Low Entropy stimuli. The continuous magenta and cyan lines represent oxyHb and deoxyHb concentrations, respectively, in response to High Entropy stimuli.

6. Discussion and Conclusions

In this study we tested how 6-month-old infants process repetition-based (ABB) regularities as compared to non-repetition-based (ABC) regularities, under low and high input entropy.

We predicted higher activation compared to baseline for both ABB and ABC sequences, which would show that 6-month-olds have the ability to process both repetition and non-repetition grammars. Indeed, we found significant activation compared to baseline for both ABB and ABC sequences collapsed over entropy conditions, in four channels (channel 6, 17 for ABB; channels 2, 17 for ABC).

We also predicted a different level of activation for the ABB sequences compared to the ABC sequences, across entropy conditions. This would show infants' ability to discriminate between repetition and non-repetition sequences, which could point to grammar learning. A channel-by-channel analysis yielded significantly different activation for ABB vs ABC only in two channels (channel 2, 20). However, contrary to our predictions, the overall analysis showed similar responses to ABB and ABC sequences in both entropy conditions. These results suggest that repetition and non-repetition grammars are processed equally, namely there are equal processing costs for repetition and non-repetition grammars. In addition, we did not find conclusive evidence that 6-month-olds discriminate between repetition and non-repetition grammars, at least not under the conditions that we tested.

Regarding the effect of input entropy on processing repetition vs non-repetition grammars, we predicted a larger difference in activation levels for the ABB vs ABC sequences in the high entropy condition compared to the lower entropy condition. However, no overall significant effect of input entropy was found across grammars, which might be due to the fact that no overall difference was found between the activation levels for ABB vs ABC sequences. In other words, a possible effect of input entropy might not be visible since it was predicted over an ABB vs ABC difference in activation, which was not found.

Notably, though, a channel-by-channel analysis showed higher activation for ABC in High Entropy than in Low Entropy in three channels (channel 5, 10, 12), higher activation for ABB in High Entropy than in Low Entropy in channel 3, higher activation for ABB in Low Entropy than in High Entropy in channel 4, and higher activation for ABC than ABB in High Entropy in channel 2. These results point to an interaction trend between grammar and input entropy, namely higher activation for repetition (ABB) grammar in low entropy, but higher activation for non-repetition (ABC) grammar in high entropy. Further research is needed in order to confirm this hypothesis.

Given that the effect over which we wanted to test the effect of input entropy was not found, i.e. an ABB vs ABC difference in activation, we suggest further research should look into a more sensitive method to capture differences between the two grammars. It might be the case that the simple interleaved block design used in this study (i.e. ABB blocks interleaved with ABC blocks) made it difficult to capture possible differences in the processing of the ABB vs ABC sequences over such a short exposure time. Specifically, the exposure time in our study was only 8 minutes compared to previous related studies that used a 15–20-minute-exposure (Bouchon et al., 2015; Gervain et al., 2008). Future fNIRS studies on this topic with repetition vs non-repetition grammars should employ alternating ABB/ABC vs non-alternating block designs, in order to increase the sensitivity of the measurements to the infant brain response.

In conclusion, at the particular input entropy values that we tested, we did not find evidence for differences in how infants process ABB vs ABC sequences. In contrast to previous findings that showed different patterns of activation for repetition vs non-repetition grammars in newborns (Bouchon et al., 2015; Gervain et al., 2008; Gervain et al., 2012), our results show similar responses for the ABB and ABC grammars. These results indicate that there might be equal processing costs for repetition and non-repetition grammars at the age of 6 months, unlike at birth. This was a surprising finding of our study, and here we discuss a few possible explanations for these findings.

In accord with the interpretation given by Gervain et al. (2012) to the finding of similar hemodynamic responses to different grammars, we interpret these results to show similar metabolic (processing) costs for encoding both repetition and non-repetition relations in 6-month-olds. These results are the first evidence pointing to a cognitive developmental change in linguistic rule learning between birth and the age of 6 months.

This change corresponds to the developmental period when infants take on some of the main learning tasks in their language acquisition endeavor, such as word learning (Tincoff & Jusczyk, 2012; Bergelson & Swingley, 2012) and grammar acquisition (Gervain et al., 2012). The ability to encode relations of non-repetition (difference), not only relations of repetition (sameness), is crucial for these learning tasks, as syllables in most words are typically different from one another and grammatical categories/rules apply to various different items. The ability to encode sequences of different items supports such learning tasks.

Independent evidence about the perception and representation of the sameness/difference relations in infants in the conceptual domain (Hochmann, Mody, & Carey, 2016) shows that 14-month-olds can complete match-to-sample and non-match-to-sample tasks. They are able to do so not only with familiar stimuli, but also with novel stimuli, thus showing generalization abilities. Authors interpreted these findings as possible evidence for infants' ability to encode relations of sameness and difference. However, in a follow-up experiment, Hochmann et al. (2016) showed that at 14 months, infants may still not be able to represent relations of difference in the conceptual domain, and instead solve the (non-)match-to-sample tasks by avoiding the sameness relation rather than encoding the relation of difference. Adding to this line of research, our study brings the first evidence that at least in the language domain the ability to perceive and process relations of difference (non-repetition) becomes available at the age of 6 months. Thus, this study contributes to a better understanding of how speech processing develops in the first year of life.

This evidence is in line with previous proposals that a primitive identity detector is in place from birth (Gervain et al., 2008), but we showed in the present study that infants' sensitivity to difference develops later on, at least by the age of 6 months. However, the exact nature of the representations of sameness/difference is still underspecified, and further research is needed in order to fully understand *how* and *why* infants firstly develop the ability to represent relations of sameness, and later on in their development the relations of difference.

Another possible logical explanation of these results, which is accounted for by our entropy model, would be the following: the *channel capacity* of infants and the exact entropy threshold that drives category formation and rule induction have not yet been determined empirically. Previous studies suggest that infants need little variability for such tasks (Gerken, 2006; Gerken, 2010; Gerken & Bollt, 2008; Gómez, 2002), but exactly where the threshold lies remains unknown. Our study is in line with this research by showing that even in the low entropy condition, 6-month-olds are able to process both ABB and ABC strings. The channel-by-channel analysis revealed sensitivity to entropy – higher activation for ABC in High Entropy vs Low Entropy in three channels, and higher activation for ABC vs ABB in High Entropy in one channel. However, we did not find an overall significant effect of input entropy. Given these findings, it is highly likely that the low entropy condition provided the necessary entropy to process both repetition and non-repetition relations, and that the higher entropy did not have a significant effect, due to the ceiling effect predicted by the *channel capacity* (Radulescu et al., 2019). According to our entropy model, *channel capacity* places a lower and an upper bound on the amount of entropy that can be encoded per unit of time. Thus, further research into this topic should expose 6-month-olds to even lower entropy in the same exposure time, in order to find the sweet-spot between *input entropy* and *channel capacity*, where the *input entropy* marks a difference in developing infants' sensitivity to repetition vs non-repetition grammars.

In this paper we implemented the same model of quantifying entropy as we did to artificial grammar learning with adults in Radulescu et al. (2019). However, given that infants' cognitive system is still under development (so their *channel capacity* is reduced) infants might be more sensitive to local statistical properties of the input rather than the entire set of items. Rather, infants might update their memory representations incrementally, in a more locally-tuned fashion (Gerken, 2010; Gerken & Quam, 2017). Indeed, as we suggested in Radulescu et al. (2019), due to lower *channel capacity*, 6-month-olds' encoding system may not be sensitive to average entropy of bigrams/trigrams over the entire set of stimuli. Thus, our findings of similar response amplitude for ABB/ABC strings in both low and high entropy conditions might be due to the particular way of calculating entropy as average over the entire set of stimuli. Future infant research on the implementation of the entropy model by Radulescu et al. (2019) should look for a method to calculate entropy on partial sets of stimuli in order to reflect an incrementally updating model.

In addition, another possibility would be that infants' sensitivity to entropy might not be fully fledged at the age of 6 moths, but only in its early stages. As we argued in Radulescu et al. (2019), sensitivity to entropy requires both encoding of sameness and difference between items. Although infants are sensitive to relations of sameness at birth (Gervain et al. 2008), their ability to process relations of difference seems to develop later in their first year of life, as suggested by the findings of our study.

Indeed, this possibility is in line with previous proposals that point to an innate advantage for representations of sameness, unlike representations of difference, which might develop at a later stage (Gervain et al. 2008; Endress et al., 2009). Not only human infants, but also non-human animals were shown to be able to perceive sameness, most likely by performing simple match computations (bees: Giurfa, Zhang, Jenett, Menzel, & Srinivasan, 2001; dolphins: Harley, Putman, & Roitblat, 2003; rats: Mumby, 2001). These animals were also shown to be able to create abstract representations of sameness, since they were able to generalize the rule to novel samples. However, the nature and the developmental trajectory of human infants' representations of sameness/difference and the related processing demands remain largely unexplored and unexplained for the first year of life.

This study brings to light the first evidence of the shift in developmental change that allows infants to represent relations of difference alongside relations of sameness at the age of six months. Thus, our study contributes to the a better understanding of the developmental trajectory and the nature of the sameness/difference representations, which underlie the building blocks of rule learning in language.

The underlying mechanisms driving the formation of sameness/difference representations (which underlie both *item-bound* and *category-based generalizations*) have remained largely underspecified, and thus heatedly debated. On the basis of studies with infants and also from similar repetition-based grammar studies with adults (Endress, Scholl, and Mehler, 2005; Endress, Dehaene-Lambertz, and Mehler, 2007), two qualitatively different mechanisms were proposed to underlie learning: abstract rule learning, based on symbolic encoding of variables (Marcus et al., 1999), and a low-level perceptual primitive ("repetition detector"), based on automatic sensitivity to repetitions (Endress et al., 2009). Whichever the mechanism, the sensitivity to repetitions seems to be in place at birth (Gervain et al., 2008). However, if learning repetition-based grammars is only supported by a perceptual primitive, then this automatic identity detector should readily enable detection of repetition in novel input as well. But this is not the case under any conditions, as Gerken (2006) and Radulescu et al. (*2019*) showed that in specific input conditions, i.e. less variability, 6-month-olds and adults, respectively, do not seem to use such a repetition detector to identify a *same-same-different* rule in novel input. Indeed, as discussed in the Introduction, Gerken (2006) found that input complexity plays a crucial role in rule learning. Moreover, several studies showed that older children and adult learners need higher variability in order to successfully generalize rules to novel input (Hudson Kam and Newport, 2009; Radulescu et al., 2019). Thus, we argue that formation of such representations recruit not only a perceptually-based repetition detector, but also a more complex mechanism that factors in a specific interaction between *input entropy* and *channel capacity,* supported by the relevant cognitive capacities (Radulescu et al., 2019).

The next interesting question would then be if the development of cognition, which is hypothesized to underlie an increase in *channel capacity,* would explain developmental changes in rule induction*.* The consequential increase in our finite time-dependent entropy-processing capacity – *channel capacity* – with age, would reduce the drive to move from a high-specificity *item-bound* form of encoding to a high-generality *category-based* form of encoding. It is a generally accepted fact that children outperform adults at language learning although their non-linguistic cognitive capacities are yet to develop. Previous studies showed that adults are more likely to reproduce the statistical properties of the input (i.e. to form high-specificity *item-bound generalizations*), while children have a higher tendency to generalize the input properties (Hudson Kam

& Newport, 2005; 2009). In accord with these researchers, we proposed that the interaction between the statistical complexity of the input – *input entropy* – and variations in *channel capacity* is key to the mechanisms of generalization (Radulescu et al., 2019). As a more fine-tuned formal approach to the *Less-is-More hypothesis* (Newport, 1990; 2016), our entropy model hypothesizes that the memory components underlying the *channel capacity* enable children to more readily (than adults) "forget" the statistical specificity of the input and generalize to novel data. Further research should investigate the exact memory components that underlie *channel capacity* and the underlying mechanism. Our entropy model makes the connection, in information-theoretic terms, between behavioral evidence found in psychological research and current hypotheses in neurobiology about the essential role of memory transience ("forgetting") in preventing overfitting to past data for the purpose of generalizing to novel environments (Richards & Frankland, 2017). This view also converges with views from neural networks research (Kumaran, Hassabis, & McClelland, 2016; LeCun, Bengio, & Hinton, 2015).

Chapter 3

Item-bound and Category-based Generalization. An Entropy Model

Radulescu, S., Giannopoulou, E., Avrutin, S., and Wijnen, F.[14]

Abstract

Language acquisition entails many learning endeavors: from segmenting speech into words, to memorizing specific items and finding statistical regularities between them (*item-bound generalization*), to forming grammatical categories to which these specific items belong, and working out relations between these categories (*category-based generalizations*). Previous research pointed to different learning mechanisms, namely a powerful domain-general statistical learning mechanism to account for segmenting speech and identifying words with their probabilistic combinations, and another more abstract algebraic rule learning mechanism for generalizing categories and relations between categories (Endress & Bonatti, 2007; 2016; Endress, Nespor, & Mehler, 2009; Marcus et al., 1999; Peña et al., 2002). Recent views converge on a single mechanism hypothesis that poses statistical learning as a powerful and economical mechanism that can account for learning both words and rules (Aslin & Newport, 2012; 2014; Christiansen & Chater, 2008; Frost & Monaghan, 2016). While supporting the single-mechanism hypothesis, this article aims at solving this debate by further extending and fine-tuning our information-theoretic model for rule induction (Radulescu et al., 2019). According to our model, the two seemingly different learning mechanisms proposed previously are actually outcomes of the same mechanism as a result of the dynamics between *input entropy* and our finite encoding capacity (*channel capacity*). Specifically, low *input entropy* facilitates not only rote memorization of the specific items in the input, but it enables *item-bound generalization,* while an *input entropy* that is higher than the available encoding capacity (*channel capacity*) drives the tendency to *gradually* move from a high-specificity *item-bound generalization* to a high-generality form of encoding, *category-based generalization*. In order to probe this hypothesis, we exposed adults to the same low and medium entropy conditions of the 3-syllable XXY grammar from Radulescu et al. (2019), and we also measured individual differences in the cognitive capacities that we

[14] This chapter is a modified version of a manuscript in preparation:
Radulescu, S., Giannopoulou, E., Avrutin, S., & Wijnen, F. (2021) Item-bound and Category-based Generalization. An Entropy Model

hypothesize to underlie the *channel capacity*, namely incidental memorization and working memory (specifically, a domain-general pattern recognition capacity that draws on working memory resources). Our findings show that indeed low *input entropy* facilitates *item-bound generalization,* not only mere memorization of specific items and of the statistical regularities present in the input. We also found that an increase in *input entropy* drives a higher tendency towards *category-based generalization.* Moreover, we found that under conditions of medium entropy, but not low entropy, learners with a low incidental memorization capacity, but a high domain-general pattern recognition capacity show a higher tendency towards *category-based generalization* than learners with a high incidental memorization capacity, but a low domain-general pattern recognition capacity, which is in line with the hypotheses of our entropy model.

1. Introduction

Besides identifying and memorizing specific items and chunks of items (e.g. words and phrases) from the input language learners need to abstract beyond specific items, to pick up relations between them, that is learn the rules of the language. From learning rules like “add *-ed*” to express past actions or “add *-ly*” to show how an action is carried out, language learners also take a further qualitative inductive step to generalizing rules to novel items, that is forming categories of items (e.g. noun, verb, adverb) and relations between these categories. For example, learners will not only memorize specific items like “walked” and “walked slowly”, but they would also infer an item-specific rule like “add a specific item *–ed* to express past actions” or “add a specific item *-ly* to show how the action takes place”. Furthermore, they will also form categories of items, like verbs, and they will also generalize further to infer that any verb can be combined with an adverb and a noun, to create a well-formed sentence. While it is a clear and undebatable fact that this rich phenomenon, named rule induction, is essential in language acquisition, it is still largely underspecified how language learners converge on these generalizations, and what drives the inductive step from memorizing specific items to rule induction.

In this paper we extend and fine-tune an information-theoretic model for rule induction that we proposed in Radulescu et al. (2019). In addition, we aim at further investigating the types of generalizations that learners converge upon, and also at probing the factors that drive the transition from item-specific rules to abstracting generalized rules that apply to novel input. In Radulescu et al. (2019), we followed suggestions from previous conceptualizations of the types of generalizations (rule induction) that learners infer (Gómez & Gerken, 2000), and we distinguished between two qualitatively different types of generalizations: *item-bound generalization* and *category-based generalization.*

More specifically, in Radulescu et al. (2019) we defined *item-bound generalization* as a relation between specific items present in the input, and identified by physical characteristics specific to a perception modality (auditory, visual, etc.). For example, a repetition relation based on physical identity, like *da-*

da (*da* is followed by the same item *da*), or an addition relation of a specific item like *-ed* or *-ly* (add *-ed* to the word *walk* or add *-ly* to the word *nice*). While *item-bound generalizations* are relations involving specific items present in a finite input set, *category-based generalizations* go beyond specific items to higher-order operations between variables (i.e. categories). For example, an *XY* relation (*Y* follows *X*), where *Y* and *X* are variables taking different items as values. Or an *XX* pattern (*X* follows *X*), where the identity relation is not based on a finite set of specific items, but it encompasses an identity relation that operates over a virtually infinite and non-specific category of items – *category X.* For example, a phonotactic generalization that allows not only for self-reduplication of specific items, e.g. *da-da, kri-kri, lo-lo* combinations, but it allows for any syllable to be followed by itself. Another example of *category-based generalizations* from natural language would be the grammatical generalization that Verb-Adverb sequences go beyond specific combinations like "walked slowly" or "talked loudly", to relations over the abstract linguistic categories, which can be construed as a relation between variable *X* that takes any verb as values and variable *Y* that takes any adverb as values.

Both young and adult learners were shown to possess domain-general learning abilities that enable both finding statistical regularities between specific items in the input (Aslin, Saffran, & Newport, 1998; Saffran, Aslin, & Newport, 1996; Thiessen & Saffran, 2007) and abstract category/rule induction (Marcus, Vijayan, Rao, & Vishton, 1999; Smith & Wonnacott, 2010; Wonnacott, 2011; Wonnacott & Newport, 2005). An already classical but still unresolved debate revolves around the question of the mechanisms that drive rule induction with its two types of generalization – *item-bound* and *category-based generalizations.* Some early studies claim that a lower-level item-specific mechanism, which relies primarily on memorization of specific items with their physical characteristics and their statistical distribution in the environment – *statistical learning* – would suffice for *item-bound generalization.* For example, infants detect patterns of specific items like phonotactic information (Chambers, Onishi, & Fisher, 2003) and word boundaries (Aslin, Saffran & Newport, 1998; Saffran et al., 1996) by statistical computations (e.g. transitional probabilities) between items they are exposed to.

On the other side, other researchers argued that, while sensitivity to statistical distribution of specific items in the input might account for item-specific learning, abstracting away from the input to form generalized rules that apply to novel instances cannot be attributed to the same limited mechanism, but to a higher-order abstract learning which operates with categories (Endress & Bonatti, 2007; 2016; Endress et al., 2009; Marcus et al., 1999; Peña, Bonatti, Nespor & Mehler, 2002). For example, Marcus et al. (1999) showed that, after exposure to strings like "leledi" and "dedeli", 7-month-olds were able to learn an abstract AAB structure to discriminate new-syllable strings with the same structure (e.g. "kokoba") from ABA-structured strings (e.g. "kobako"). Hence, the authors argued that infants possess an abstract symbolic ('algebraic') system that operates over variables and relations between them. Specifically, Marcus et al. (1999) proposed that learners have two qualitatively different mechanisms at hand: *statistical learning* for tracking and computing statistical distributions of specific items (i.e. *item-bound generalization* in our terminology), and *abstract rule learning* for operations beyond specific items with variables (i.e. *category-based generalization* in our terminology).

More recent views in computational models of rule learning proposed that a dichotomy between the two mechanisms would not suffice to explain the rich generalization phenomena observed in both young and adult learners, and that learners might actually add generalization to statistical learning in phonotactic rule induction (Adriaans & Kager, 2010), or statistical inference might be performed over built-in rule-based representations (Frank & Tenenbaum, 2011). Another view suggests that one single mechanism – *statistical learning* – accounts for both types of generalization (Aslin & Newport, 2012; 2014; Frost & Monaghan, 2016). Aslin & Newport (2014) argue in favor of a *single-mechanism hypothesis* based on the finding and proposal that learners show a *gradient of generalization* depending on the statistical properties of the input they are exposed to (Gerken, 2006; Reeder, Newport & Aslin, 2013). More specifically, Reeder et al. (2013) and Aslin & Newport (2013; 2014) claim that learners show a different pattern of learning depending on the consistency of contexts for the items in the input (e.g. the preceding and/or following items), such that learners either converge on abstract rule learning, when many items occur interchangeably in the same context (i.e. contexts are largely shared), or they withhold generalization, so that there is only surface statistical learning, when the contexts apply only to specific items (i.e. there are consistent gaps in the shared contexts). Nevertheless, although recent views converge on the *single-mechanism hypothesis* which underlies an apparent *gradient of generalization*, *how* and *why* this gradual mechanism outputs two qualitatively different forms of generalization – *item-bound and category-based generalization* – remains largely underspecified.

Two other lines of related research investigated the factors that drive rule induction and showed that both item-specific learning and abstract generalization are modulated by either external factors (i.e. input variability – Gerken, 2006; 2010; Gerken & Bollt, 2008; Gómez, 2002; Reeder et al., 2013), or internal factors (i.e. young learners' limited memory capacity as compared to adult learners – Hudson Kam & Chang, 2009; Hudson Kam & Newport, 2009; Newport, 1990). In the remaining part of this introduction we will briefly review these previous findings and hypotheses related to the external and internal factors that are claimed to drive generalization, with its two types – *item-bound* vs *category-based generalization.* Next, we will introduce our information-theoretic entropy model which supports the *single-mechanism hypothesis* and puts together both external and internal factors in order to give a consistent account for both forms of generalization (Radulescu et al., 2019).

Regarding infant rule induction, Gerken (2006) initiated a line of research that looked into the external factor of input variability, and showed that 9-month-old infants exposed to four AAB strings (e.g. *leledi*) ending in different syllables (*je/li/di/we*) generalized to novel strings with an AAB structure (e.g. *kokoba,* which was not heard in the familiarization), i.e. in our terminology, they made *category-based generalizations*. However, the other group of infants, who were exposed to four AAB strings ending only in *di,* did not make a broad generalization – AAB, but only a narrow generalization – *"every string ends in di"* – that is, in our terminology, they only made an *item-bound generalization.* In a subsequent study (Gerken, 2010), 9-month-olds were presented with the same "ends in *di*" condition from Gerken (2006), with the crucial difference that at the end of the familiarization with numerous strings ending in *di* infants heard only three strings ending in "*je/we/li*". This little input variability was sufficient to drive the broader AAB generalization. These studies together with others (Gerken & Bollt, 2008; Gómez, 2002) show that infant rule induction is driven by input variability, and that infants need little variability in order to move from *item-bound* to *category-based generalization.*

Adult rule induction studies have also shown that input variability is the main factor that drives generalization. For example, Reeder et al. (2013) familiarized adults with a *(Q)AXB(R)* grammar (where each letter stands for a three-word category), and asked whether learners generalize *X* as a category of items rather than just memorize the exact strings, when exposed to different subsets of strings (i.e. with different number of combinations of words from each category) generated by the grammar. Participants were asked to give grammaticality judgements on the novel (withheld) grammatical strings, as well as on ungrammatical strings (*(Q)AXA(R)* or *(Q)BXB(R)* strings). The results showed that rich combinations between words from all categories (i.e. a high input variability, in our terminology) drove high tendency towards generalizing *X* as a category (i.e. *category-based generalization,* in our terminology).

In Radulescu et al. (2019), we took a step further from the previous studies by showing that the driving factor for rule induction is not just mere input variability in the sense of number of items in the input, but it is a particular pattern of input variability captured by *input entropy* (as a function between number of items and their probability distribution – Shannon (1948)). More specifically, we exposed adults to six versions of a 3-syllable XXY artificial grammar (where each letter stands for a category of items), with different *input entropy* in each version. In the test phase learners were asked grammaticality judgements on XXY strings with familiar syllables and with new syllables. The results showed that the tendency to accept new XXY strings as grammatical increased *gradually* as a function of increasing *input entropy,* in accord with the information-theoretic entropy model proposed in Radulescu et al. (2019). These findings brought more granular evidence for the previously proposed *gradient of generalization* (Aslin & Newport, 2012; 2014) and better specified the positive effect of input variability on rule induction which was found in previous studies (Gerken, 2006; 2010; Gerken & Bollt, 2008; Gómez, 2002; Reeder et al., 2013).

Another line of research investigated the internal factors (i.e. learners' cognitive capacities) that drive rule induction. The classical *Less-is-More* hypothesis (Newport, 1990; 2016) and subsequent related studies (Hudson Kam & Chang, 2009; Hudson Kam & Newport, 2005; 2009) propose and show evidence for children's higher tendency to move away from the statistical specificity of the input and overregularize the input as compared to adults, who remember and rather stick to the probability distributions specific to the input. However, adults were also shown to generalize away from the probability distribution of the items in the input, when exposed to a memory overloading and noisier input, i.e. high variability (Hudson Kam & Newport, 2009; Hudson Kam & Chang, 2009).

Under the *Less-is-More* hypothesis, children's tendency to generalize was assumed to be driven by their incomplete cognitive development (maturational constraints – Newport 1990, 2016), more specifically by their limited memory capacity (children's overall lower memory capacity – Cowan, 1997; Gathercole, 1998). Thus, the findings and claims of this line of research can also be interpreted to bring evidence in favor of an internal factor, i.e. cognitive capacities, that drives rule induction.

However, *how* and *why* cognitive limitations, such as memory limitations, could have an impact on generalization remains largely underspecified. In fact, some studies (Perfors, 2012) did not find any evidence in favor of an effect on generalization of overloading the working memory with concurrent tasks, or any evidence for an effect on generalization of individual differences in working memory (measured independently in a complex span task, which is a widely used task to measure working memory capacity – Conway et al., 2005; Unsworth, Spillers, & Brewer, 2009).

As discussed above, previous research found some evidence for a link between the capacity for rote memorization and a *gradient of generalization*. While supporting the previously proposed *single-mechanism hypothesis* with a *gradient of generalization* (Aslin & Newport, 2014), in order to bridge the gap between previous lines of research, we aim at better specifying and investigating this gradual mechanism, and the dynamics between the internal and external factors. Specifically, we deem rule induction to be a phased mechanism, that starts out by rote memorization of the items and their probability distribution (or statistical regularities of the input), and *gradually* moves to *item-bound generalization* and, eventually, to *category-based generalization.* In Radulescu et al. (2019) we proposed that the underlying processes (i.e. the learning mechanisms proposed in previous research – *statistical learning* and *abstract rule learning)* should be conceptualized separately from their outcomes, that is from the resulting types of generalization (*item-bound generalizations* and *category-based generalizations*). This distinction allowed for the main hypothesis of our information-theoretic model to be formulated. In short, and simplifying for now, our entropy model hypothesizes that *item-bound generalization* and *category-based generalization* are not independent mechanisms, but they are outcomes of the same phased encoding mechanism which *gradually* moves from memorized combinations of items to *item-bound generalizations*, and, eventually, to *category-based generalizations*.

Rule induction is hypothesized to be an encoding mechanism driven by the interaction between an external factor – the statistical properties of the input, i.e. *input entropy* – and an internal factor – the brain's ability to encode the input under conditions of finite encoding capacity (i.e. *channel capacity*). *Channel capacity* quantifies the maximum rate of information encoding, i.e. amount of entropy that can be encoded per unit of time. We define our encoding capacity as *channel capacity,* in information-theoretic terms, which is the finite rate of information encoding (entropy per unit of time).

Specifically, if *input entropy* is lower than the available *channel capacity,* the input can be encoded using rote memorization and/or high-specificity *item-bound generalization*, while an increase in *input entropy* gradually shapes *item-bound generalization* into *category-based generalization*, in order to avoid exceeding the *channel capacity*. We also hypothesize that the *channel capacity* associated with rule induction is supported by certain cognitive capacities, e.g. memory capacity, in psychological terms.

As we argued in Radulescu et al. (2019), our entropy model specifies *how* and *why* this gradient emerges: the gradual – *bit by bit* – accumulation of entropy per unit of time up to the upper bound placed by the available *channel capacity* drives a change in encoding in order to allow for higher *input entropy* to be encoded more efficiently (i.e. with the least loss of information). In order to test the gradual transition from rote memorization to *item-bound* to *category-based generalization* hypothesized by our entropy model, in Radulescu et al. (2019), we exposed adults to six increasingly entropic versions of a 3-syllable XXY artificial grammar. After the exposure phase, participants were asked for grammaticality judgements on XXY strings with experienced (familiar) and new syllables. As a test for gradual transition from *item-bound generalization* to *category-based generalization,* we expected a gradually increasing tendency to accept not only Familiar-syllable XXY strings as grammatical, but also New-syllable XXY strings. Indeed, as predicted, results showed gradual acceptance of New-syllable XXY strings as grammatical, with a constantly high acceptance of Familiar-syllable XXY strings, as a function of increasing *input entropy*. These results showed that learners had an increasing tendency to form a *same-same-different* generalization (XXY structure) not only with specific syllables (i.e. familiar syllables experienced in the familiarization), but crucially with any syllables (i.e. novel syllables never heard in the familiarization). However, since low entropy allows for easy memorization of the input, one might argue that in the low entropy conditions high endorsement of Familiar-syllable XXY strings (basically the same strings from familiarization) might have been supported by rote memorization of the specific strings, not necessarily by *item-bound generalization* (i.e. *same-same-different* structure with specific familiar syllables).

Thus, in the current study, we aim at further investigating the type of encoding that low input entropy facilitates, and at better specifying the conditions under which *item-bound generalization* is formed and gradually shaped into *category-based generalization.* To this end, we exposed adults to the lowest entropy condition and a middle entropy condition from Radulescu et al. (2019), but, crucially, we swapped the Familiar-syllable XXY test strings with Familiar-syllable YYX strings. In other words, instead of the exact familiarization strings (e.g. *da:-da:-li)* participants were tested on strings like *li-li-da:*. The rationale was that, if indeed low *input entropy* facilitates *item-bound generalization*, learners would not just memorize the strings as such, with the exact sequence of items. Rather, they would encode the input as a *same-same-different* structure with familiar items, which would allow them to accept strings with the *same-same-different* structure, but with switched over syllables (e.g. *li-li-da:*).

This kind of encoding is a case of *item-bound generalization,* as defined above: *item-bound generalizations* are relations involving specific items present in a finite input set. More specifically, learners not only retain by rote memorization a repetition pattern based on item-specific positional information (according to the probability distribution of items in the input, i.e. a transitional probability of 1), that is only those specific items that reduplicate themselves in the input can be duplicated. Learners infer a self-duplication rule in the first positions of the triplets also for those specific items from the input that do not show a reduplication pattern in the input. For example, input strings show reduplication of *da:,* but not of *li* (like *da:-da:-li*)*,* however learners accept *li-li-da:* as grammatical. So, if low *input entropy* facilitates *item-bound generalization*, not just mere memorization of the strings, learners are expected to accept YYX strings with familiar syllables, that is strings with a *same-same-different* structure with switched over syllables heard in the familiarization. Moreover, since an increase in *input entropy* is hypothesized to drive the transition from *item-bound generalization* to *category-based generalization,* learners exposed to the medium entropy version of the XXY grammar would not only accept the YYX strings (i.e. a *same-same-different* structure with familiar syllables), but crucially they would show a higher tendency than learners in the low entropy condition to accept the *same-same-different* structure with new syllables as well (i.e. New-syllable XXY strings).

In this study, we also measured, in independent tests, learners' individual cognitive capacities that our entropy model hypothesizes to underlie *channel capacity*, namely incidental memory capacity and a domain-general pattern recognition capacity that draws on working memory resources (Radulescu et al., 2019), in order to look into the effect of these cognitive capacities on rule induction.

2. An entropy model for rule induction

In Radulescu et al. (2019), we proposed an information-theoretic entropy model that hypothesizes rule induction to be driven by a single encoding mechanism. More precisely, the model predicts that variations in the ratio between *input entropy* and our capacity to encode a finite amount of entropy per second (i.e. *channel capacity*) gives birth to rule induction with its two flavors – *item-bound generalization* and *category-based generalization*. In short, our entropy model poses that an increase in *input entropy* drives a gradual transition from rote memorization of specific configurations of items in the input, to *item-bound* to *category-based generalization*. This transition happens as a means to develop a more efficient encoding method (i.e. with the least loss of information), which would avoid exceeding the *channel capacity*. We use the concepts and formulas for entropy and channel capacity as they were introduced and mathematically demonstrated by Shannon (1948).

For a random variable *X*, with *n* values {$x_1, x_2 \ldots x_n$}, Shannon's entropy (Shannon, 1948), denoted by *H(X)*, is defined as:

$H(X) = -\sum_{i=1}^{n} p(x_i) log p(x_i)$ [15];

where *p(xi)* is the occurrence probability of x_i. H measures the entropy per symbol produced by a source of input, relative to all the possible symbols (values) contained by the set (Shannon, 1948).

In Radulescu et al. (2019), we proposed an innovative method to calculate entropy of an artificial grammar based Shannon's formula and on a similar calculation method proposed by Pothos (2010) for finite-state grammars. In the experiments reported in Radulescu et al. (2019), adults were exposed to a 3-syllable XXY artificial grammar, in six experimental conditions with increasing input entropy. Results showed that when input entropy increases, the tendency to move from *item-bound* to *category-based generalization* increases gradually.

Here we further extend and elaborate on the predictions related to the effect of the *input entropy* on rule induction, which were proposed in Radulescu et al. (2019):

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 rote memorization of the specific items and their exact configuration in the input, to a high-specificity form of encoding (*item-bound generalization)* to a high-generality encoding (*category-based generalization)*. This gradual transition is driven by the interaction between *input entropy* and the finite encoding capacity of the brain (*channel capacity)*, as follows:

a. Low *input entropy* allows for rote memorization of items and chunks of items with their probability distribution in the input (i.e. probability matching[16]). However, low *input entropy* – that is below or matches the *channel capacity* – promotes also an encoding method – *item-bound generalization* – that not only matches the exact exposure probabilities of the items. Rather, the encoding is a *generalization* beyond the particular probability distribution or the transitional probabilities of the experienced input, but, crucially, restricted to specific items (e.g. experienced stimuli).

The specific items present in the input are encoded with their uniquely-identifying (acoustic, phonological, phonotactic, prosodic, etc.) features, but their probability distribution is flexible and can be generalized. An example of such a generalization in natural language would be an addition rule of a specific item *-ly* or *-ed* to words. For instance, a learner exposed to an input

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

[16] *Probability matching* is a term dubbed in the literature on regularization patterns (e.g. Hudson Kam & Newport, 2009) and refers to learning and preserving the specific probability distribution of items in the input rather than systematically imposing rules on the variation of the input.

sentence like "Maria is nice and speaks clearly", would find it acceptable to say "Maria speaks nice*ly*"; also a learner exposed to "John ask*ed* Maria to call him." would accept "Maria call*ed* John".[17]

This form of encoding is an intermediate step on the encoding continuum from specificity to generality, which preserves the input structure, but, crucially, only with the specific items present in the input, hence the name *item-bound generalization.*

Thus, if the *input entropy* is lower than the *channel capacity*, the input can be encoded by a high-fidelity *item-bound generalization*, and transmitted through the channel at the available channel rate – *channel capacity* (i.e. amount of entropy per unit of time).

b. If the *input entropy* is higher than the finite *channel capacity* of the encoding system, that is the *input entropy per second* supported by the channel, it is possible to find a proper method that encodes more information (entropy), but the rate of information encoding cannot exceed the available *channel capacity* (Theorem 11 – Shannon, 1948).

More precisely, when the input entropy per second is higher than the available *channel capacity*, the *item-bound generalization* becomes inefficient and prone to many errors, due to a high number of items, encoded with high-specificity, and a complex probability distribution. Thence, the information cannot be encoded with a high-fidelity method (i.e. *item-bound generalization*), because this encoding method leads to high loss of information. As we argued in Radulescu et al. (2019), it is essentially the finite *channel capacity* which "forces" re-structuring of the information in order to *gradually* – bit by bit – shape the *item-bound generalization* into another less specific, and thus more general form of encoding.

As we argued in Radulescu et al. (2019), information is re-structured by (unconsciously) re-observing the item-specific features and their configurations, and un-binding items from previously formed structures, and by identifying similarities and differences in order to compress the information. Information compression entails *gradually* reducing the number of specific features encoded with individual items, i.e. erasing or "forgetting" insignificant differences, which are low probability features. Erasing – "forgetting" – the specific features, i.e. differences, results in more similarities being highlighted between items, such that they are grouped in variables (i.e. categories) based on non-specific shared features. This marks the birth of a new higher-level form of encoding (i.e. *category-based generalization*), which supports encoding of higher *input entropy* using the same available rate of information encoding per unit of time, i.e. *channel capacity.*

[17] The tendency to over-produce specific frequent items (e.g. determiners) regardless of their probability distribution in the input, i.e. *regularization*, which was found in previous studies (Hudson Kam & Chang, 2009; Hudson Kam & Newport, 2005; 2009), is an example of *item-bound generalization,* in our terminology.

Thus, the increase in *input entropy* and the interaction with the finite *channel capacity* drives re-structuring of the information, in accord with Dynamic Systems Theory (Prigogine & Stengers, 1984; Schneider & Sagan, 2005) invoked also in studies of other cognitive mechanisms – e.g. Stephen, Dixon, and Isenhower, 2009) for the purpose of adapting to noisier environments (i.e. increasingly entropic environments).

2. As outlined in the hypotheses above, we use *channel capacity* to model the finite encoding capacity of the learning system in information-theoretic terms, i.e. at the computational level, in the sense of Marr (1982). In order to formulate a hypothesis regarding the underlying cognitive capacities (i.e. at the algorithmic level), we follow experimental evidence from the *Less-is-More hypothesis* line of research (Hudson Kam & Newport, 2005; Hudson Kam & Newport, 2009; Newport, 1990; 2016) and also classical and more recent theoretical models of memory and attention (Baddeley, 2000, 2007, 2012; Baddeley, Eysenck, & Anderson, 2015; Cowan, 1988, 1995, 1999, 2005; Miller, 1956; Oberauer & Hein, 2012). Thus, we hypothesize that the cognitive capacities underlying *channel capacity,* hence having an effect on rule induction, are a capacity for rote (unintentional) memorization of specific items (Baddeley et al., 2015), and also the attentionally-controlled regions of activated long-term memory, i.e. working memory (Cowan, 1988, 1995, 1999, 2005; Oberauer and Hein, 2012).

The rote unintentional memorization capacity (Baddeley et al., 2015) is hypothesized to have a negative effect on the transition from *item-bound* to *category-based generalization*, since a strong memory capacity for specific items and their probability configuration would support a higher *input entropy* to be encoded per unit of time, (i.e. a higher *channel capacity,* in computational terms). The rationale for proposing the unintentional (incidental) memorization capacity would be that, according to our entropy model, rule induction is a natural automatic mechanism driven by the brain's sensitivity to entropy, not a rational mechanism that consciously looks for and remembers specific items and probability distributions in the environment.

Conversely, the attentionally-controlled regions of activated long-term memory, that is working memory (WM) capacity (Cowan, 1988, 1995, 1999, 2005; Oberauer and Hein, 2012) is hypothesized to have a positive effect on the transition from *item-bound* to *category-based generalization*, since it provides the information storage and online time-dependent processing resources that are needed for faster and efficient re-structuring of the information into new configurations, as described above to be the trademark of shaping the *item-bound generalization* into *category-based generalization.*

Generally, we think that linguistic rule induction is supported by a domain-general WM capacity, rather than language-specific algebraic rule learning as proposed by early prominent research (Marcus et al., 1999). However, in the current study we explore specific possible memory components and WM-correlated abilities that we deem to be directly involved in linguistic rule induction, in order to add to the more general storage and retrieval components tested in previous studies under the *Less-is-More hypothesis* –

Hudson Kam & Chang, 2009; Perfors, 2012). Thus, we predict that one of the components underlying *channel capacity* in linguistic rule induction is a domain-general pattern recognition capacity. This hypothesis is based on the rationale that a rule induction task can be intuitively envisaged as a task of finding patterns/rules in the input.

A candidate test for a domain-general pattern recognition capacity is the Raven's Standard Progressive Matrices (RAVENS test – Raven, Raven, & Court, 2000), since it was shown to be based on rule induction (Carpenter, Just & Shell, 1990; Little, Lewandowsky, & Griffiths, 2012). Even though there is no identity (Conway, Kane, & Engle, 2003) or causality relationship (Burgoyne, Hambrick, & Altmann, 2019) between the ability measured by RAVENS test and the WM capacity, high positive correlations were found between measures of WM capacity and tests for this domain-general pattern-recognition capacity (like RAVENS – e.g. Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Little, Lewandowsky and Craig, 2014; Dehn, 2017).

To summarize, our entropy model predicts a *single mechanism* (of statistical nature), which outputs a *gradual* transition from rote memorization (of items and chunks) to *item-bound generalization* to *category-based generalization*, in accord with the previously suggested *gradient of generalization* (Aslin & Newport, 2014). Furthermore, it refines and extends this proposal, by explaining *how* and *why* this gradual process happens. Entropy captures the continuum from specificity to generality, and quantifies it in bits of information. As we argued in Radulescu et al. (2019), sensitivity to entropy under conditions of finite *channel capacity* entails a sensitivity to a specific pattern of variability in the input given by the degree of sameness/difference between items and also their probability distribution, which assigns significance (or stability) to specific combinations of items.

The number of differences encoded with each item gives the degree of specificity of the encoding (i.e. *item-bound* specificity). High number of differences, which is quantified by *entropy*, defines a high-fidelity encoding (i.e. *item-bound generalization*). If the upper bound placed by the *channel capacity* on the number of bits encoded per second is reached, a reduction (or "gradual forgetting") of the differences is triggered. As a result, the similarities between items gain a higher weight, which drives an automatic gradual grouping of items under the same category. Gradually, the specificity decreases, while the degree of generality increases with each bit of information reencoded.

3. Experiment: Design and Rationale

The goal of this study is to further probe the effect of *input entropy* on rule induction as hypothesized by our entropy model, by specifically looking at the kind of encoding method that low entropy facilitates, and at the transition from *item-bound* to *category-based generalization* from low to medium entropy in an XXY grammar. In our previous experiments reported in Radulescu et al. (2019), we exposed adults to six increasingly entropic versions of a 3-syllable XXY grammar (e.g. *ke:ke:my, da:da:li*), namely in six experimental conditions with increasing input entropy – 2.8, 3.5, 4, 4.2, 4.58, 4.8 bits – and in all entropy conditions we found the same high tendency to accept Familiar-syllable XXY strings (i.e. the same XXY strings they were familiarized with). However, participants showed a gradually increasing percentage of correct acceptance of New-syllable XXY test strings (i.e. *same-same-different* structure with new syllables, e.g. *dy-dy-ta:*), with a very low at-chance acceptance rate in the lowest entropy condition, and increasing *gradually* in the medium to high entropy conditions. Altogether, we interpreted these results to show an increasing tendency to move from *item-bound generalization* towards *category-based generalization* as a function of increasing *input entropy* (Radulescu et al., 2019).

However, those results did not show very clearly what type of encoding was actually formed in the low entropy conditions and the medium entropy conditions. Since low-to-medium entropy allows for easy rote memorization of the specific XXY strings from familiarization, the high tendency to accept Familiar-syllable XXY strings might have been supported by a strong memory trace of the exact XXY strings from the familiarization, and not necessarily by an *item-bound generalization* (i.e. *same-same-different* rule with familiar syllables). Also, in the low entropy condition, *category-based generalization* did not support this performance either since it was not developed (i.e. at-chance acceptance of New-syllable XXY strings).

Hence, the question that we address in this paper is whether in conditions of low-to-medium entropy learners make use of plain rote memorization, or they actually form *item-bound generalizations* as hypothesized by our model. In other words, does low entropy facilitate *item-bound generalization* and not just rote memorization*?* Also does an increase in *input entropy* from low to medium entropy drive the transition from *item-bound generalization* to *category-based generalization*?

In order to answer these questions, we exposed participants to the same low and medium entropy conditions from Experiment 2 in Radulescu et al. (2019), but crucially we changed one type of test items in order to have a more in-depth understanding of the encoding developed in a low entropy environment as compared to a medium entropy environment.

More precisely, participants were exposed (aurally) to an artificial XXY grammar using the same stimuli as those used in Radulescu et al. (2019) in the lowest entropy condition (here Low Entropy condition – 2.8 bits) and one of the medium entropy conditions (here Medium Entropy condition – 4.25 bits). In the test phase, just as in the design by Radulescu et al. (2019), participants in both conditions were presented with the same grammaticality judgement task, where they had to answer a yes/no question to indicate whether the test strings could be possible in the familiarization language. Crucially, in this study, we changed one of the four types of test strings from Radulescu et al. (2019). While in Radulescu et al. (2019), one of the test types was Familiar-syllable XXY strings, which were the exact same XXY strings heard in the familiarization (e.g. *ke:-ke:-my, da:-da:-li*), in this study, we replaced the Familiar-syllable XXY test items with Familiar-syllable YYX test items, that is participants were familiarized with XXY strings like *ke:-ke:-my, da:-da:-li*, but were tested on strings like *my-my-ke:, li-li-da:*. Therefore, we tested the learners on the four types of test strings presented below, and we had the following predictions for the performance in the Low Entropy condition vs Medium Entropy condition.

Familiar-syllable YYX (*same-same-different* structure with familiar but switched over X-syllables and Y-syllables) – correct answer: yes – accept. According to the hypotheses of the entropy model, low entropy facilitates *item-bound generalization,* thus we expect learners in the Low Entropy condition to accept this type of test strings as possible in their familiarization language. Similarly, participants in the Medium Entropy condition are expected to accept this type of test items, either by having encoded the input as *same-same-different* structure with familiar syllables (*item-bound generalization*), or as *same-same-different* structure with any syllables (*category-based generalization*).

New-syllable XXY (*same-same-different* structure with new X-syllables and Y-syllables) – correct answer: yes – accept. This test type probes whether learners' *item-bound generalization* was shaped into *category-based generalization,* which would allow them to accept XXY strings with new syllables (i.e. *same-same-different* structure regardless of familiar or new syllables). According to the hypotheses of the entropy model, we expect differences between the entropy groups, since low entropy is assumed to not be higher than the *channel capacity*, while medium entropy would drive a higher tendency to move from *item-bound* to *category-based generalization.* Thus, we expect significantly higher acceptance rates of these strings in the Medium Entropy condition as compared to the Low Entropy condition. However, as we argued in Radulescu et al. (2019), absolute mean acceptance rate of this type of strings should not be interpreted as direct evidence for *category-based generalization*. This mean should be compared to the mean acceptance rate of Familiar-syllable YYX strings: the smaller the difference of the mean acceptance rate (i.e. the effect size) between New-syllable XXY strings and Familiar-syllable YYX strings is, the more likely it is that learners encoded the input as having a *same-same-different* structure regardless of new or familiar syllables. Having encoded the input as *same-same-different* structure with any items means that learners would not make a distinction between the New-syllable XXY strings and Familiar-syllable YYX, so they have moved from *item-bound* to *category-based generalization*.

Familiar-syllable X_1X_2Y (structure of three different but familiar syllables) – correct answer: no – reject. According to the entropy model, participants are expected to reject this type of strings, either by having encoded the input as *item-bound* or *category-based generalizations*. Specifically, participants in the Low Entropy condition are expected to reject this type of strings, since they are expected to encode the input as *item-bound generalizations,* that is *same-same-different* structure with familiar syllables, which would highlight mismatches with a *different-different-different* structure with familiar syllables. Learners in the Medium Entropy condition are also expected to reject these strings, due to a higher tendency towards encoding the input as *category-based generalization,* which would highlight a violation of the *same-same-different* structure with any syllable.

New-syllable X_1X_2Y (structure of three different and new syllables) – correct answer: no – reject. In both conditions, participants are expected to reject this type of strings, either by having encoded the input as *item-bound* or *category-based generalizations*.

In addition to probing the effect of *input entropy* on the transition from *item-bound* to *category-based generalization*, as presented above, this study also looked into the effect of the cognitive capacities hypothesized by our entropy model to underlie the *channel capacity*: unintentional memory capacity and a domain-general pattern-recognition capacity, as argued in the previous section.

Thus, we tested each participant on three individual tests: a Forward Digit Span task, which is a measure of explicit short-term memory (Baddeley et al., 2015), an incidental memorization task, which measures implicit short-term memory capacity (Baddeley et al., 2015), and Raven's Standard Progressive Matrices (Raven et al., 2000), a standardized test based on visual pattern-recognition (Caroll, 1993; Conway et al., 2002). Thus, according to the hypotheses of our entropy model, we predicted a positive effect of RAVENS test on the tendency to move from an *item-bound* to a *category-based generalization*, and a negative effect of the explicit/incidental memory tests on the same transition from one type of encoding to the other.

Moreover, in order to further probe the effect of individual differences in the cognitive capacities that we hypothesize to play a role in rule induction, we planned an analysis based on a post-hoc split of the participants in four groups, depending on their scores on the incidental memory test and the RAVENS test. Specifically, we planned to group the participants based on a median split of the scores on the incidental memory test into Low Memory and High Memory groups, and based on a median split of the scores on the RAVENS test into Low RAVENS and High RAVENS groups. Then, each participant would be assigned to one of the four possible resulting Incidental Memorization Task - RAVENS groups (ImtRAV groups): 1. Low Memory – Low RAVENS (participants with lower than median score on Incidental Memorization test, but lower than median score on RAVENS test), 2. Low Memory – High RAVENS (participants with lower than median score on Incidental Memorization test, but higher than median score on RAVENS test), 3. High Memory – Low RAVENS (participants with higher than median score on Incidental Memorization test, but lower than median score on RAVENS test), 4. High Memory – High RAVENS (participants with higher than median score on Incidental Memorization test, but higher than median score on RAVENS test). According to the hypotheses of our entropy model, we predicted that participants with low incidental memory capacity, but high visual pattern-recognition capacity (Group 2. Low Memory – High RAVENS) would show the highest tendency towards *category-based generalization*. Conversely, participants with high incidental memory capacity, but low visual pattern-recognition capacity (Group 3. High Memory – Low RAVENS) would show the lowest tendency towards *category-based generalization.*

4. Methods

4.1 Participants

97 healthy, non-dyslexic Dutch speaking participants (74 females, 23 males, age 18-46, M=23.7) were assigned to either the Low Entropy condition (46 participants) or the Medium Entropy condition (51 participants). Four more participants were tested in the Low Entropy condition, but excluded because of self-reported general knowledge about artificial grammar research. Only healthy participants that had no known language, reading or hearing impairment or attention deficit were included. They all signed a form of consent and were paid for their participation.

4.2 Tasks and materials

Task 1: XXY grammar

Familiarization stimuli. In the Low Entropy condition, participants listened to the same XXY artificial grammar (*X* and *Y* stand for non-overlapping sets of syllables) used in the low entropy condition of Experiment 2 from Radulescu et al. (2019), while in the Medium Entropy condition, participants listened to the XXY stimuli used in the medium entropy condition of Experiment 2 from Radulescu et al. (2019). All strings of the grammar in both conditions are three-syllable strings of the language with a *same-same-different* structure: each string has two identical syllables (XX) followed by another different syllable (Y), e.g. *ke:ke:my, da:da:li*. All syllables are natural Dutch syllables having the same structure, i.e. a consonant followed by a long vowel. For the Low Entropy condition, 7 X-syllables and 7 Y-syllables were used to generate seven strings (see Appendix A for complete stimulus set). All seven strings were repeated four times (7 strings * 4 = 28 strings) in each familiarization phase (there were three familiarization phases, each consisted of the same 28 strings), so that the entropy was the same in each familiarization phase – 2.8 bits. For the Medium Entropy condition, we used 14 X-syllables and 14 Y-syllables (we added another set of 7 different X-syllables and 7 different Y-syllables to those used in the Low Entropy condition), and each syllable was repeated 2 times. We spliced the syllables into 28 XXY strings, which were used in each of the three familiarization phases, so that the entropy was the same in each familiarization phase – 4.25 bits. The order of presentation of the strings was randomized for every participant. We used the same method for the entropy calculations as in Radulescu et al. (2019), which is a fine-tuned extension of a related entropy calculation method proposed by Pothos (2010) for finite state grammars (see Table 1 below for complete entropy calculations).

Medium Entropy	Low Entropy

H[bX]=H[14] = 3.8
H[XX] = H[14]= 2.8
H[XY] = H[28] = 4.8
H[Ye] = H[14] = 3.8
H[bXX] = H[14] = 3.8
H[XXY] = H[XYe]= H[28] = 4.8
H[bigram] = 4.05
H[trigram] = 4.46
H[total] = $\frac{\text{H[bigram]+H[trigram]}}{2}$ **= 4.25**
	H[bX]=H[7] = -Σ[0.143*log0.143] = 2.8
	H[XX] = H[7]= 2.8
	H[XY] = H[7] = 2.8
	H[Ye] = H[7] = 2.8
	H[bXX] = H[7] = 2.8
	H[XXY] = H[XYe]= H[7] = 2.8
	H[bigram] = 2.8
	H[trigram] = 2.8
	H[total] = $\frac{\text{H[bigram]+H[trigram]}}{2}$ **= 2.8**

Table 1. Entropy values for Low Entropy and Medium Entropy conditions. Taken from Experiment 2, Radulescu et al. (2019)

Test stimuli.

The three familiarization phases were interleaved with three (quick) intermediate test phases and a final (longer) test phase. Each intermediate test phase included four test strings, one of each of the four types presented in the previous section. The final test had eight test strings (two of each type). Thus, in total, there were (4+4+4+8=) 20 test strings (see Appendix A for the complete set of test stimuli). Accuracy scores for the learning of the XXY grammar were measured as correct acceptance of strings with the *same-same-different* structure (Familiar-syllable YYX and New-syllable XXY strings), and correct rejection of strings that deviate from the *same-same-different* structure (Familiar-syllable X_1X_2Y and New-syllable X_1X_2Y strings).

Task 2: Forward Digit Span

Participants were instructed to listen to series of digits presented aurally, and they were told in advance that it was a memory test. Participants had a short trial phase to become familiar with the task, and then the actual test began: participants listened to the digits (audio files of auditory recordings of the digits), and they were asked to enter the digits they heard in the same order. We modified the design of the classical Forward Digit Span task, such that participants did not have any physical keyboard, but a row with buttons for each digit was displayed in a line on the screen only in the moment when they were asked to enter the digits, and disappeared during the listening phases. In order to enter the digits, the participants had to click on the buttons displayed on the screen, not to use a physical keypad on a keyboard. This modification was intended to prevent participants from creating a visual pattern on the physical keypad of the keyboard while listening to the digits. The row of digits disappeared from the screen when the next series of digits was presented aurally. The task was progressively difficult, starting out with a series of 3 digits and ending with a series of 12 digits (in total 24 trials – 2 for each series of digits). After 2 consecutive mistakes the task ended automatically. The score was the highest span achieved by the participant, i.e. the largest series for which both trials were correctly completed.

Task 3: Incidental Memorization Test

In this task, participants were not told in advance that this was a memory test. Instead, they were only told that they would listen to words from another forgotten language. Participants listened to 30 bisyllabic nonsense words resembling Dutch phonology and phonotactics. They were instructed to imagine what the word might have meant in the forgotten language and to pick a category (flower, animal, or tool) for each word they heard, based on what the word sounded like to them. They had 3 seconds to choose a category for each word, by pressing the button for flowers, animals, or tools.

After the listening/categorization phase was over, a surprise instruction appeared on the screen, informing the participants that they would be given a memory test, which would check whether they remembered the words they categorized during the previous phase. They were instructed to indicate whether they heard the word previously, by clicking a yes/no button on the screen. The memorization test consisted of 13 targets and 13 foils.

Task 4: RAVENS

Participants were administered 5 sets of matrices, with 12 matrices per set to resolve. Each matrix consisted of nine visual patterns (of which one pattern is missing) arranged in a particular order in accordance with some underlying rules. Participants have to find the missing pattern for each of the sixty matrices in a multiple-choice task.

4.3 Procedure

Participants were tested in a sound-proof booth, and they completed the tasks in the order presented above.

For Task 1 – XXY grammar, the participants were told that they would listen to a "forgotten language" that would not resemble any language they might be familiar with, but which had its own rules and grammar. The instructions informed participants that the language had more words than the ones played in the familiarization phases. They were also explained that there would be three familiarization phases interleaved with three intermediate tests and a final (longer) test phase, which were meant to check what they had noticed about the language. They were asked to judge, whether the test words could be possible in the language that they listened to, by pressing a Yes/No button. This task lasted around 5 minutes.

After the first task, participants were given the instructions for the Forward Digit Span, namely they were explicitly instructed that it was a memory test. They were asked to listen attentively to streams of digits, which they would have to recall in the same order. This task lasted around 5 minutes.

The third task was the Incidental Memorization task, for which participants were told in advance that they would have to listen to words from another “forgotten language”. Their task would be to imagine what the words might have meant in the forgotten language, based on how the words sounded like to them. Importantly, participants were not told in advance that a memory test would follow. This task lasted about 7 minutes.

Lastly, participants were asked to perform the RAVENS matrices test, as a paper-and-pen task. The standard RAVENS task allows participants to spend 50 minutes in total, but, after running a pilot testing, we modified the task to allow participants only a shorter amount of time (35 minutes) to complete the task, in order to avoid an overall time-consuming and exhausting experimental session. The experimenter would walk in 20 and 30 minutes after participants started the session, to announce the remaining time. The entire experiment lasted about one hour.

4.4 Data scoring and analysis

We recorded all the yes/no answers for Task 1 and coded them as correct or incorrect as per the criteria presented for each type of strings in Section 3 above. From all the 20 correct/incorrect answers for each participant, a proportion of correct answers was calculated per each type of test item. Instead of directly analyzing proportions, we performed an empirical logarithmic transformation, in order to analyze the data using a linear regression model.

For the Forward Digit Span task, the standard scoring method was used, that is the measured highest span of each participant was recorded as one data point per participant. In the Incidental Memorization Task, all correct/incorrect answers were recoded into hits and false alarms, which were used to calculate a *d’* value for each participant. For the RAVENS test, we used the standard scoring method, that is all correct answers to all sets of matrices were counted, and the count was transformed into age-corrected percentiles using the standardized RAVENS tables.

5. Results

Figure 1 presents the mean correct acceptance rate (proportion of correct acceptances per group) for Familiar-syllable YYX strings and New-syllable XXY strings, across the two conditions (Medium Entropy and Low Entropy). The mean correct acceptance rate in the Medium Entropy condition for Familiar-syllable YYX strings was *M* = .9 (*SD* = .15), and for New-syllable XXY strings it was *M* = .82 (*SD* = .21). The mean rate of correct acceptance in Low Entropy condition for Familiar-syllable YYX strings was *M* = .96 (*SD* = .11), and for New-syllable XXY strings it was *M* = .62 (*SD* = .32).

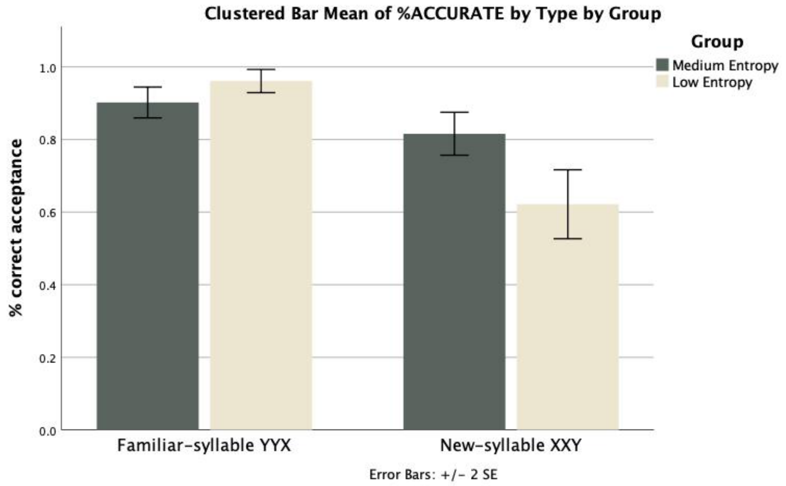


Figure 1. Mean rate of correct acceptance for Familiar-syllable YYX and New-syllable XXY strings in both conditions: Medium Entropy and Low Entropy. Error bars show standard error of the mean.

Similarly, Figure 2 shows the mean correct rejection rate (proportion of correct rejections per group) for Familiar-syllable X_1X_2Y strings and New-syllable X_1X_2Y strings, across the Low Entropy and Medium Entropy conditions. In the Medium Entropy condition, the mean correct rejection rate for Familiar-syllable X_1X_2Y strings was *M* = .84 (*SD* = .3) and for New-syllable X_1X_2Y strings it was *M* = .9 (*SD* = .17). In the Low Entropy condition, the mean correct rejection rate for Familiar-syllable X_1X_2Y strings was *M* = .75 (*SD* = .37), and for New-syllable X_1X_2Y strings it was *M* = .91 (*SD* = .16).

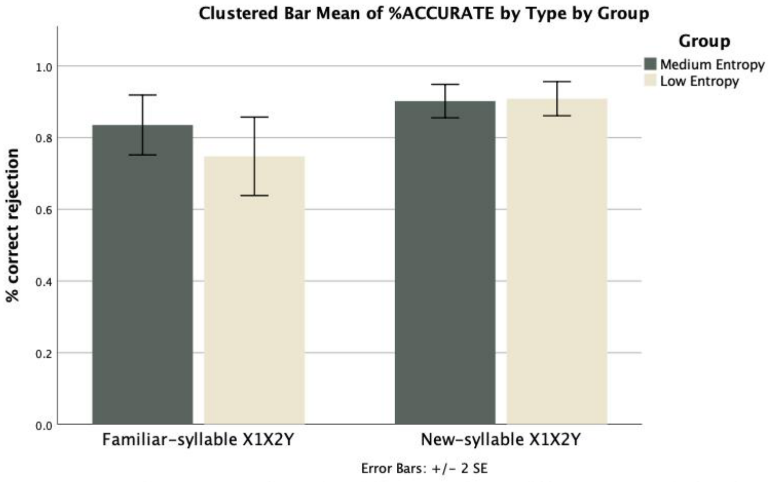


Figure 2. Mean rate of correct rejection for Familiar-syllable X1X2Y and New-syllable X1X2Y strings in both conditions: Medium Entropy and Low Entropy. Error bars show standard error of the mean.

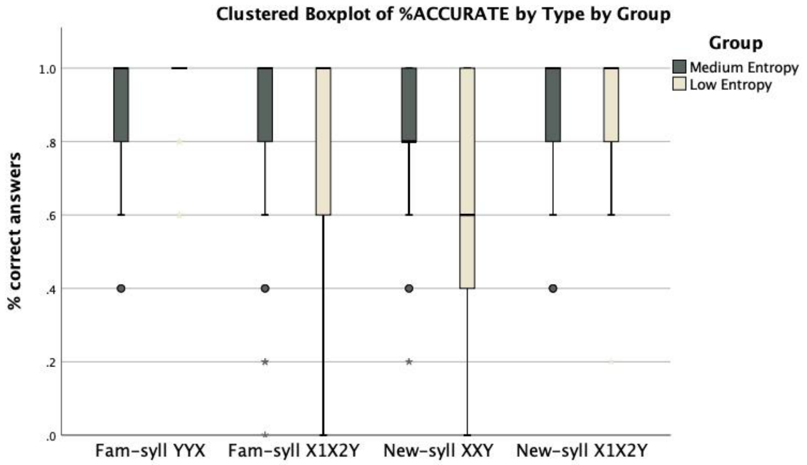


Figure 3. On the X-axis, the four types of test items: Familiar–syllable YYX; Familiar–syllable X1X2Y; New–syllable XXY; New syllable X1X2Y. On the Y-axis the mean rate of correct answers: correct acceptance for Familiar–syllable YYX and New–syllable XXY and correct rejection for X1X2Y strings (with familiar or new syllables), in both groups: Medium Entropy and Low Entropy.

Figure 3 shows the distribution of individual mean rates for each test type in each experimental condition, Low Entropy and Medium Entropy.

In order to probe the effect of *input entropy* on rule induction, we compared the performance in the two conditions (Medium Entropy and Low Entropy groups) in a general linear mixed effects analysis of the relationship between Accuracy (correct acceptance of the grammatical test items and correct rejection of the ungrammatical ones) and Type of Test (Familiar-syllable YYX, New-Syllable XXY, Familiar-syllable X1X2Y, New-Syllable X1X2Y), Group (Medium Entropy and Low Entropy), as well as Group x Type of Test interaction. Therefore, as dependent variable we entered log-transformed Accuracy score into the model. As fixed effects we entered Type of Test, Group and Group x Type of Test interaction. The scores for Forward Digit Span, Incidental Memorization Task and RAVENS tests were entered one by one as covariates in the model, and their interactions with Group was also entered one by one in the model. As random effect we had an intercept for subjects. 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. The model reported here is the best fitting model, both in terms of the model's accuracy in predicting the observed data, and in terms of AIC (Akaike Information Criterion).

We found a significant Group x Type interaction (F(7, 377) = 8.761, *p* < .001), a non-significant Group x Incidental Memorization Task interaction (F(2, 377) = 1.498, *p* = .225), and a significant main effect of RAVENS (F(1, 377) = 3.890, *p* = .049).[18]

[18] None of the other factors, interactions between factors or covariates had a significant effect, and since they did not improve the model they were removed from the final model reported here.

Pairwise comparisons of the Estimated Marginal Means (adjusted to the mean values of the covariates in the model, i.e. Incidental Memorization Task = 1.575, RAVENS = 75) revealed a significant difference between Groups (Medium Entropy and Low Entropy groups) for the Familiar-syllable X1X2Y (M = .168, SE = .066, F(1, 377) = 6.412, *p* = .012) and the New-syllable XXY (M = .21, SE = .066, F(1, 377) = 10.037, *p* = .002). For the other two Types of test, pairwise comparisons of the Estimated Marginal Means adjusted for the same level of the covariates revealed a non-significant difference between Groups (Medium Entropy and Low Entropy groups): Familiar-syllable YYX (M = - .036, SE = .066, F(1, 377) = .300, *p* = .584) and New-syllable X1X2Y (M = - .007, SE = .066, F(1, 377) = .013, *p* = .911).

Further, Cohen's effect size value for the mean difference in correct answers between the Medium Entropy and the Low Entropy groups was d = - .45, with an effect size correlation *r* = - .22 (Familiar-syllable YYX), d = .27, *r* = .13 (Familiar-syllable X1X2Y), d = .74, *r* = .35 (New-syllable XXY) and d = - .06, *r* = - .03 (New-syllable X1X2Y). The effect size for the difference between acceptance of Familiar-syllable YYX vs. New-syllable XXY was higher in the Low Entropy group (Diff of Means = .34, *d* = 1.42, *r* = .58) compared to the same difference in the Medium Entropy group (Diff of Means = .08, *d* = .44, *r* = .21).

In order to further look into the effect of individual differences in Incidental Memorization (Imt) and RAVENS scores, we grouped the participants post-hoc, separately in each Entropy Condition, into four groups (ImtRAV groups) based on a median split of their individual scores on these tests, as follows: 1. Low Memory – Low RAVENS (participants with lower than median score on Incidental Memorization test, but lower than median score on RAVENS test), 2. Low Memory – High RAVENS (participants with lower than median score on Incidental Memorization test, but higher than median score on RAVENS test), 3. High Memory – Low RAVENS (participants with higher than median score on Incidental Memorization test, but lower than median score on RAVENS test), 4. High Memory – High RAVENS (participants with higher than median score on Incidental Memorization test, but higher than median score on RAVENS test). In order to probe the effect of individual differences in Incidental Memorization and RAVENS scores on rule induction, we compared the performance between the four post-hoc ImtRAV groups (Low Memory – Low RAVENS, Low Memory – High RAVENS, High Memory – Low RAVENS, High Memory – High RAVENS), in each Entropy Condition separately, in planned ANOVAs of the relationship between Accuracy (correct acceptance of the grammatical test items and correct rejection of the ungrammatical ones) and ImtRAV Group. Thus, as dependent variable we entered log-transformed Accuracy scores into the analysis.

In the Medium Entropy condition, we found a significant effect of ImtRav group (F(3, 200) = 7.110, *p* < .001) on the Accuracy scores across all test types. Multiple Bonferroni-corrected post-hoc comparisons showed that Group 1. Low Memory – Low RAVENS had significantly higher accuracy scores than Group 3. High Memory – Low RAVENS (M = .15, SE = .05, *p* = .04), Group 2. Low Memory – High RAVENS had significantly higher accuracy scores than Group 3. High Memory – Low RAVENS (M = .22, SE = .05, *p* < .001), and Group 4. High Memory

– High RAVENS had significantly higher accuracy scores than Group 3. High Memory – Low RAVENS (M = .2, SE = .05, *p* < .001). Specifically, by Test Type, we found a significant difference between ImtRAV groups on Accuracy scores on the Familiar-syllable X1X2Y test strings (F(3, 47) = 2.929, *p* = .043), with Group 4. High Memory – High RAVENS performing significantly better than Group 3. High Memory – Low RAVENS (M = .46, SE = .17, *p* = .05). We also found a significant difference between ImtRAV groups on Accuracy scores on the New-syllable XXY strings (F(3, 47) = 6.033, *p* = .001), with Group 1. Low Memory – Low RAVENS performing significantly better than Group 3. High Memory – Low RAVENS (M = .2, SE = .06, *p* = .017), Group 2. Low Memory – High RAVENS performing significantly better than Group 3. High Memory – Low RAVENS (M = .25, SE = .06, *p* = .001) and Group 4. High Memory – High RAVENS performing significantly better than Group 3. High Memory – Low RAVENS (M = .18, SE = .06, *p* = .013). We did not find a significant effect of ImtRAV group on Accuracy scores for the Familiar-syllable YYX test strings (F(3, 47) = 2.038, *p* = .12) or for the New-syllable X1X2Y test strings (F(3, 47) = 1.259, *p* = .29).

In the Low Entropy condition, we did not find a significant effect of ImtRav group (F(3, 180) = .338, *p* = .79) on the Accuracy scores across all test types.

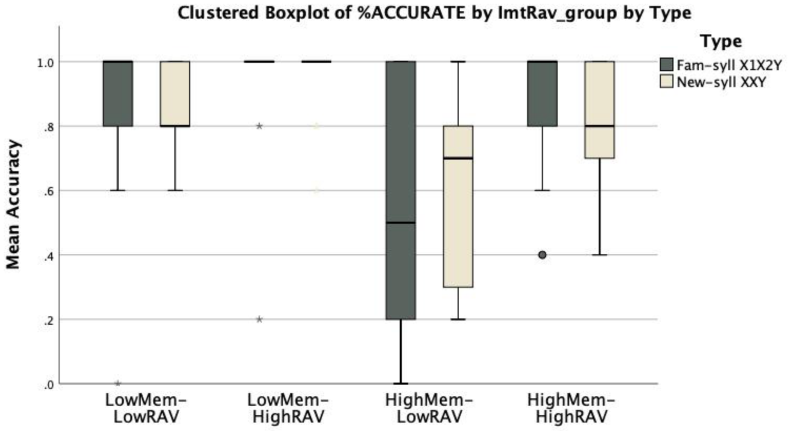


Figure 4. The distribution of individual mean accuracy rates per test type in each group: LowMem–LowRAV (n=40), LowMem–HighRAV (n=52), HighMem–LowRAV (n=32), HighMem–HighRAV (n=80), in the Medium Entropy Condition.

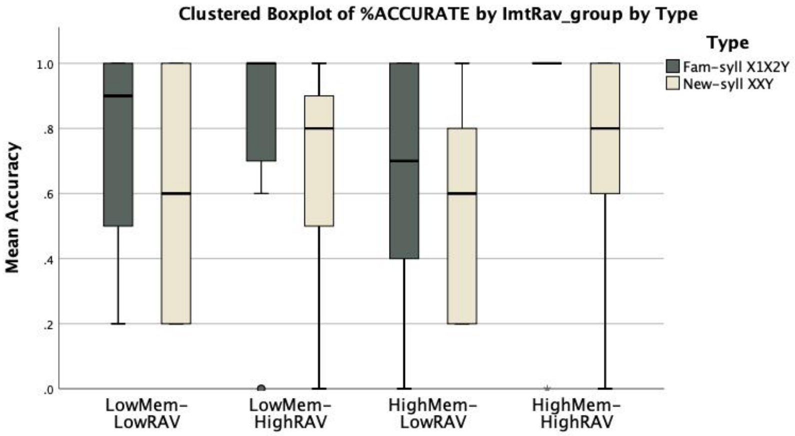


Figure 5. The distribution of individual mean accuracy rates per test type in each group: LowMem–LowRAV (n=32), LowMem–HighRAV (n=44), HighMem–LowRAV (n=56), HighMem–HighRAV (n=52), in the Low Entropy Condition.

Figures 4 and 5 show the distribution of individual mean accuracy rates for the test types[19] in each ImtRAV group, for each experimental condition, Low Entropy and Medium Entropy.

6. Discussion

The goal of this experiment was to probe the effect of *input entropy* on the transition from *item-bound* to *category-based generalization*, and to further investigate the type of generalization that low *input entropy* facilitates. The results showed that when exposed to a 3-syllable XXY grammar, with strings like *ke:-ke:-my, da:-da:-li*, in both a low entropy and a medium entropy condition, adults have a similar high tendency in both entropy conditions to accept as grammatical familiar-syllable YYX strings, that is strings with familiar but switched over syllables, e.g. *my-my-ke:, li-li-da:*. These results show that learners in both low and medium entropy conditions did not only encode the input by rote memorization of the familiarization strings, but they took a step further and encoded the input strings as having a *same-same-different* structure with familiar syllables. We interpret these results as evidence in favor of our entropy model, which hypothesizes that low *input entropy* facilitates *item-bound generalization,* not only rote memorization of the items and of the surface statistical regularities (e.g. transitional probabilities) between the items displayed in the input.

[19] For conciseness, we only show boxplots for the test types where we found significant differences in the Medium Entropy condition, i.e. Familiar-syllable X1X2Y and New-syllable XXY. For visual comparison reasons, we show boxplots for the same test types in the Low Entropy condition, although we did not find significant differences in this condition.

Moreover, we found that learners exposed to the medium entropy version of the language showed a higher tendency to accept new XXY strings as grammatical than learners in the low entropy condition. In addition, the difference between acceptance of Familiar-syllable YYX strings compared to New-syllable XXY (i.e. the effect size) was higher in the low entropy group than the same difference in the medium entropy group. This shows that learners in the medium entropy condition had a higher tendency than those in the low entropy condition to encode the input as having a *same-same-different* structure regardless of familiar or new syllables, which means they abstracted away from the specific items in the input and their configuration (i.e. probability matching), and encoded the input as relations over variables. This finding supports the hypothesis of our entropy model: an increase in *input entropy* from low to medium entropy drives the transition from *item-bound* to *category-based generalization.*

Furthermore, we found a higher tendency to correctly reject the ungrammatical Familiar-syllable X1X2Y strings in the medium entropy condition than in the low entropy condition. This type of strings consisting of three different but familiar syllables poses a challenge to the learner, in that a strong memory trace of the familiar syllables might incorrectly lead to acceptance of these strings, if a *same-same-different* structure is not strongly encoded, which could trigger rejection as an XXY-rule violation. Since medium entropy drove strong development of a *same-same-different* structure with any syllable, not only the familiar ones, i.e. *category-based generalization*, this form of encoding supports rejection of the Familiar-syllable X1X2Y more strongly than the *item-bound generalization* driven by the low entropy language version. These results are in line with our findings in Radulescu et al. (2019), where in the medium entropy condition correct rejection of Familiar-syllable X1X2Y strings was supported by *category-based generalization,* since the rejection rate of Familiar-syllable X1X2Y was just as high as the acceptance rate of New-syllable XXY.

Regarding the effect of individual differences in the cognitive capacities hypothesized to underlie the *channel capacity,* we found a significant positive effect of the individual scores in the domain-general pattern-recognition RAVENS test, in both entropy conditions. This result is in line with the hypothesis of our entropy model regarding a positive effect of working memory, in particular of a domain-general pattern recognition ability (RAVENS test). These findings support the hypothesis that rule induction in language is supported by a domain-general pattern-recognition capacity, which was shown to be highly correlated with working memory capacity (Little, Lewandowsky and Craig, 2014; Conway et al., 2002), based on the fact that it draws on the attentionally-modulated storage and processing resources that help keeping goal-relevant information active in the face of concurrent processing (i.e. the trademark of working memory – Baddeley et al., 2015; Conway et al., 2002). Therefore, a high domain-general pattern recognition capacity supports a higher tendency to move from *item-bound* to *category-based generalization* under conditions of low-to-medium *input entropy*.

Moreover, in the medium entropy condition, but not in the low entropy condition, we found significant differences in overall accuracy scores (i.e. correct acceptance of *same-same-different* strings, and correct rejection of X1X2Y strings) between post-hoc groups of learners with low or high incidental memorization and low or high domain-general pattern recognition capacity. Specifically, these group differences were significant in the correct acceptance of the new XXY strings: the Low Memory – Low RAVENS group had higher accuracy scores than the High Memory – Low RAVENS group, which shows that when it comes to rule induction, individuals with a low domain-general pattern recognition capacity might benefit from a lower incidental memorization capacity. Moreover, the Low Memory – High RAVENS showed significantly higher accuracy scores than the High Memory – Low RAVENS group, and also the highest accuracy scores among all the groups, which shows that individuals with a low incidental memorization capacity, but a high domain-general pattern recognition capacity have a higher tendency towards *category-based generalization* than the individuals with high incidental memorization capacity and high domain-general pattern recognition capacity. Finally, the High Memory – High RAVENS group had significantly higher accuracy scores than the High Memory – Low RAVENS group, which shows that individuals with high memorization capacity might benefit from a high domain-general pattern recognition capacity in their tendency towards *category-based generalization*. In short, a high domain-general pattern-recognition capacity drives better rule induction, however, when the pattern-recognition capacity is low, a low incidental memory capacity drives higher tendency towards *category-based generalization* than a high incidental memory capacity. This shows that, in line with our entropy model hypothesis regarding cognitive capacities underlying *channel capacity,* generally a combination of low incidental memorization capacity and a high working memory capacity, specifically a high domain-general pattern-recognition capacity, drive the tendency to move from *item-bound* to *category-based generalization,* under conditions of medium *input entropy*.

The fact that in the low input entropy condition we did not find significant differences between the combinations of these cognitive capacities could have a possible logical explanation, under the hypotheses of our entropy model. The low entropy we employed was so low that individual differences in incidental memorization would not make a difference. Specifically, even low memory learners could easily remember the seven strings (repeated four times) so that their low incidental memorization capacity would not give them an advantage towards *category-based generalization.* However, the domain-general pattern-recognition capacity as a main effect played a positive role in driving *category-based generalization* regardless of *input entropy*.

Our findings together with our entropy model contribute to the hotly-debated topic of the learning mechanisms underlying statistical regularities and rule induction. Here we will briefly revive and discuss the previous arguments in order to integrate our entropy model and findings in the general debate on statistical learning vs rule induction. Previously, it was claimed that computing statistical regularities displayed by the input, e.g. transitional probabilities between experienced items, and abstracting away from the specific items in the input to make higher-order generalizations, e.g. category formation and syntactic structure, are qualitatively distinct mechanisms (Endress & Bonatti, 2007; 2016; Endress et al., 2009; Marcus et al., 1999; Peña et al., 2002). While it is widely accepted and established by mounting evidence that statistical learning underlies learning tasks based on computations of probabilistic distributions of specific items, such as phonotactic information (Chambers et al., 2003), speech segmentation (Aslin et al. 1998; Saffran et al., 1996) and learning co-occurrence dependencies between items in sequences (Gómez, 2002; Lany & Gómez, 2008; Lany, Gómez & Gerken, 2007), the sophisticated rule induction mechanism that enables category formation and generalization to novel instances remains largely underspecified, and hence hotly debated (Aslin & Newport, 2012; 2014; Christiansen & Chater, 2008; Christiansen & Curtin, 1999; Frost & Monaghan, 2016; Radulescu et al., 2019).

Generally the arguments in favor of different mechanisms are built around two main assumptions: one related to a very basic (narrow) interpretation of sensitivity to statistical learning, i.e. probabilistic computations that can only apply to experienced (familiar) stimuli, but not to unexperienced (novel) ones, and another assumption related to an apparent temporal distinction between the two mechanisms, i.e. lack of simultaneity.

Having initiated the first argument, Marcus et al. (1999) claimed that a basic statistical learning mechanism which relies on computations of probabilities between experienced stimuli cannot account for generalizing to novel instances, for which another abstract algebraic-rule mechanism would be necessary. A myriad of neural network studies followed the algebraic-rule proposal suggesting mainly the following counterargument: if neural network models, which are seen as an implementation of statistical learning, lacking any symbolic (algebraic) representations of rules, can capture the regularities in the input (i.e. repetition-based structure in the case of Marcus et al.'s stimuli), thus mirroring the human performance, then this could be taken as evidence that symbolic/algebraic representations might not be necessary and that a statistical mechanism can in principle account for rule learning in humans (Altmann, 2002; Altmann & Dienes, 1999; Christiansen & Curtin, 1999; Gasser & Colunga, 2000; Seidenberg & Elman, 1999; Sirois et al., 2000; for a recent review of such neural networks and symbolic models, see Alhama & Zuidema, 2019).

While having thoroughly informed the field on the type of computations that might theoretically be involved in rule induction, neural network based approaches cannot be considered direct evidence regarding the mechanisms employed by the human brain. Also, concrete empirical evidence that the living brain uses these mechanisms, however biologically plausible, remains elusive. Moreover, neural networks might have to at least be combined with symbol-manipulation mechanisms in order to reach human-level productivity (Marcus, 2001; 2013). Even the latest state-of-the-art deep learning algorithms, while showing generalization capabilities, are still far less efficient than humans in learning complex rules (Lake, Salakhutdinov, & Tenenbaum, 2015; Lake, Ullman, Tenenbaum, & Gershman, 2017; Marcus, 2018). In their defense, neural network model studies (Christiansen & Curtin, 1999; Christiansen, Conway, & Curtin, 2000; Seidenberg & Elman, 1999; Sirois et al., 2000) argue that even earlier models of simple recurrent networks were able to replicate the generalization behavior from Marcus el al's study (1999). The same was argued regarding abstract recurrent networks, which have a built-in short-memory and an identity detector as a prior mechanism (Dominey & Ramus, 2000). More advanced deep learning neural networks (e.g. convolutional neural networks – LeCun, 1989; LeCun, Bengio, & Hinton, 2015) are built on a simplicity principle, such that prior built-in knowledge is purposefully minimized in order to create simple models of the data. This is achieved by constraining the bits of information represented by the synaptic weights, which in turn leads to better generalization ability (LeCun, Denker, & Solla, 1989; MacKay, 2003). Thus, we think that the latest deep learning neural networks harness the strengths of a property that mirrors a similar design feature of the biological memory system: preventing overfitting to past data enables better generalization (Moscovitch, Cabeza, Winocur, & Nadel, 2016; Richards & Frankland, 2017).

In light of the findings of the present study (and of Radulescu et al., 2019), we challenge the proposal that our mind is innately endowed with a symbol-manipulation mechanism (Marcus, 2001): why is it that the use of such a mechanism depends on other factors, namely, the input entropy and certain cognitive capacities? As shown in previous studies discussed in the Introduction of this article, and as per the findings of this study, learners do not employ the abstract rule mechanism unless the input entropy reaches a certain threshold or unless the learner's incidental memorization and domain-general pattern recognition have a certain capacity. Theoretically, it might be the case that only under certain conditions will the abstract mechanism be triggered, that is only when necessary for efficiency purposes (i.e. extracting abstract rules requires usage of extra resources, such that it is triggered only when extracting rules is computationally more efficient than memorizing all the items and their statistical regularities). Nevertheless, the findings of a *gradual* tendency to generalize as a function of increasing input entropy (as shown in this study and in Radulescu et al., 2019) would be quite difficult to account for by a built-in symbol-manipulation mechanism theory, and they challenge the plausibility of multiple mechanisms coexisting.

Assuming the other argument – the temporally distinct mechanisms argument – Endress and Bonatti (2007) proposed a *More-than-One-Mechanism hypothesis* to account for what they claimed to be two different types of learning mechanisms, namely, a statistically-driven mechanism that accounts for learning co-occurrences (by computing transitional probabilities) between specific items in the input, and another mechanism capable of extracting structural regularities, like classes of words and rules. Their hypothesis assumes that the structure-extracting mechanism outputs the structural information faster than the statistical mechanism, which needs time (i.e. exposure to repeated exemplars) in order to strengthen the memory traces of the specific items.

Thus, they suggest speed of representation formation to be a test for the type of mechanism, and they claim that learning classes of items (i.e. *category-*

based generalization, in our terminology) and learning associations between items (i.e. *item-bound generalization*, in our terminology) are different mechanisms specifically because they are temporally distinct processes. Specifically, Endress and Bonatti (2007) exposed adults to streams of nine three-syllable A_iXC_i words that followed a non-adjacent dependency pattern, where A_i always paired with C_i. They found participants' tendency to make class-based generalization decreased linearly with longer exposure (i.e. tendency to accept $A_iX'C_j$ class-words, where any *A* syllable could be paired with any *C* syllable, not a strict item-based relation between a particular A_i syllable and a particular C_i syllable; the middle syllable *X'* had never occurred in the middle position in the familiarization, but was one of the *A* or *C* familiarization syllables). Based on these results, the authors conclude that learners possess two qualitatively different mechanisms: statistical mechanisms for computing statistical regularities (e.g. transitional probabilities in speech segmentation) and generalization mechanisms, which are responsible for grammatical generalizations (Endress & Bonatti, 2007; 2016). The temporally distinct mechanisms argument (Endress & Bonatti, 2007; 2016; Peña et al., 2002) was subsequently challenged by Frost & Monaghan (2016), who showed that speech segmentation and generalization of non-adjacencies from continuous speech occur simultaneously, and thus they proposed a single statistical learning mechanism to account for both processes in the absence of evidence to the contrary.

In any case, we think that in principle a temporal distinction argument does not necessarily hold to support a multiple-mechanism hypothesis, because what researchers conceptualize as two qualitatively different types of generalization might be outcomes of a single phased mechanism, under different conditions (Radulescu et al., 2019). Just like physically different and temporally successive states of water in nature as iced water, liquid water and evaporated water do not imply different mechanisms underlying the phase transition, in this case heating is the single mechanism that underlies the time-dependent qualitative change in state (which is driven, neither accidentally nor randomly, by increasing entropy).

Specifically about certain grammars, e.g. repetition-based grammars like the XXY grammar used in this study or the ABB grammar used by Marcus et al. (1999), the multiple-mechanism hypothesis takes yet another assumption into account. Endress and colleagues (Endress, Dehaene-Lambertz, & Mehler, 2007; Endress et al., 2009) challenge Marcus et al.'s (1999) proposal that repetition-based grammars are learned by extracting variables and relations between them, and they argue that in order to learn a repetition-based grammar of the ABB type a low-level perceptual primitive, a "repetition detector", would suffice. Indeed a low-level perceptual identity detector ("repetition detector") is in place from birth (Gervain, Berent, & Werker, 2012; Gervain et al., 2008) and it might aid learning of repetition-based grammars. Such a perceptual primitive would supposedly suffice to find identity of items in the input, regardless of familiar or new stimuli: just as *le-le-di* can be recognized by a repetition-detector primitive as an instance of a *same-same-different* pattern with familiar syllables, the new *ko-ko-ba* can be recognized as an instance of the same pattern, without the need for abstract variables to be extracted.

However, if learning of a repetition-based XXY grammar only requires a repetition detector, why do learners in low entropy conditions do not apply the repetition-detector to new XXY strings (as we found in the present study and in Radulescu et al., 2019)? Also, why do learners exposed to a low entropy XXY grammar do not equally apply the same repetition-detection mechanism to both Familiar-syllable YYX strings and New-syllable XXY strings? The findings of the present study showed that in the low entropy condition learners show higher tendency to accept Familiar-syllable YYX strings than they do with New-syllable XXY strings, although both types of strings display the *same-same-different* pattern immediately recognizable by a repetition-detector.

Based on our findings, it seems plausible to conclude that learners of an XXY grammar not only apply a perceptual identity primitive on the surface item-specific features. Rather, they encode the actual items themselves (as hinted at by Aslin & Newport (2014)) and they keep track (in the working memory) of the familiar items (experienced syllables in the familiarization). Depending on the entropy of the set of items tracked, they either encode relations (rules) only between familiar syllables (that is *item-bound generalization,* in our terminology), or they generalize these rules also to novel syllables, under higher input entropy (*category-based generalization*). Moreover, learners not only retain an identity pattern based on item-specific positional information (according to the probability distribution of the items in the input), that is only those specific items that replicate themselves in the input can be duplicated. Learners infer a self-duplication rule in the first positions of the triplets also for those items that do not show a reduplication pattern in the input. This acceptance shows they encoded the input by *item-bound generalization*, that is a *same-same-different* generalization, but only with the familiar items, not generalized to novel items. When the input entropy is higher though, learners not only detect the *same-same-different* pattern between experienced stimuli, but show also a higher tendency then in the low entropy to generalize the rule to novel items, showing thus *category-based generalization.*

7. Conclusion

In this study we further examined the effect of *input entropy* on rule induction as hypothesized by our information-theoretic entropy model (Radulescu et al., 2019). According to our model, an *input entropy* that is lower than the available *channel capacity* facilitates high-specificity *item-bound generalizations*, while a higher *input entropy* than the *channel capacity* drives a *gradual* transition to high-generality *category-based generalization.* While our previous results showed the gradual transition towards *category-based generalization* as a function of increasing *input entropy* (Radulescu et al., 2019), here we further investigated and better specified the type of generalization that low entropy facilitates. To this end, we further probed the kind of regularities that learners infer under low *input entropy* as compared to medium entropy.

Specifically, we exposed adults to the lowest and the medium entropy versions of the 3-syllable XXY grammar from Radulescu et al. (2019). We asked whether low *input entropy* indeed facilitates *item-bound generalization*, not only mere memorization of the familiarized strings and the statistical regularities in the input. We also asked whether an increase up to a medium entropy drives a higher tendency towards *category-based generalization.* To address the first question, we exposed adults to 3-syllable XXY strings (e.g. *da:-da:-li*), and asked them for grammaticality judgements on YYX strings with familiar syllables (i.e. strings with a *same-same-different* structure with familiar, but switched over syllables – e.g. *li-li-da:*). To address the second question, we asked for grammaticality judgements on XXY strings with new syllables, that is strings with a *same-same-different* structure, but with syllables that never occurred in the familiarization.

We hypothesized that, if learners accept familiar-syllable YYX strings, it means they encoded the input as *item-bound generalizations*, while a higher tendency towards accepting new-syllable XXY strings also, besides the familiar-syllable YYX strings, shows they moved towards *category-based generalization.* Indeed, as expected, the results showed a high acceptance of familiar-syllable YYX strings both in the low entropy condition and in the medium entropy condition. However, in the medium entropy condition there was a higher acceptance rate of the new-syllable XXY strings as compared to the low entropy condition. Taken together, these results bring further evidence in favor of our entropy model, which hypothesizes that low entropy indeed facilitates *item-bound generalization*, not only rote memorization of the items and of their statistical regularities present in the input (e.g. transitional probabilities), while an increase in *input entropy* drives the transition from *item-bound* to *category-based generalization.* In terms of cognitive capacities underlying *channel capacity*, we found evidence that, generally, learners with a high domain-general pattern-recognition capacity and a low incidental memorization capacity have a higher tendency to move from *item-bound* to *category-based generalization* compared to learners with a low domain-general pattern-recognition capacity and a high incidental memorization capacity.

Given the hypothesis of a *gradual* transition from *item-bound* to *category-based generalization* made by our entropy model, supported by our previous findings (Radulescu et al., 2019) and the findings of this study, one might ask about the nature of representations as a continuum from *item-bound* to *category-based generalization.* More specifically, how could the representations be envisaged as graded on the continuum from *item-bound* to *category-based generalization*? Previous studies that proposed a *gradient of generalization* (Aslin & Newport, 2012; 2014) left this question unanswered, mainly because their proposal, although dubbed *gradient of generalization*, only focused on two categorical outcomes: learners either restrict generalization or they generalize, depending on the consistency of distributional contexts for the items in the input. Unlike previous proposals, we conceptualize the two flavors of rule induction – *item-bound* and *category-based generalization* – as two qualitatively different outcomes of a *gradual* encoding mechanism, however we do not think that there is a dichotomy clearly represented with a clear shift from one to another.

As we showed in Radulescu et al. (2019) and further extended in the current study, learners *gradually* accept the grammatical strings as a function of increasing *input entropy*, moving from acceptance of the *same-same-different* structure with familiar syllables only, but not with new syllables, to gradually higher acceptance of this structure both with familiar and new syllables*.* This behavioral tendency could be interpreted in at least two ways: either it reflects learners' gradually increasing levels of confidence in their hypothesis about the structure of the input, or it might reflect a sort of a dynamic and fuzzy representation of the input structure, which is updated gradually – *bit by bit* – as the learner's environment becomes increasingly entropic. Although from the kind of evidence we bring in the current study and in Radulescu et al. (2019), it is not possible to establish with certainty which interpretation fits the data better, we suggest that it might not be needed to choose between the two. Specifically, the latter interpretation reflects the nature of rule induction as a *gradual* encoding which moves on the continuum from specificity to generality depending on the *input entropy* and the available rate of information encoding (i.e. bits/second – *channel capacity*)*.* As a natural result of the fuzzy nature of the representation on the specificity-to-generality continuum, learner's confidence in their hypothesis about the input structure gradually changes, as well.

Another question which logically follows from the previous one would be: what exactly is the mechanism that drives the dynamics of this fuzzy representation? In other words, what is the exact *gradual* mechanism of rule induction under conditions of increasing *input entropy* and our cognitive capacities? While unraveling the sophisticated mechanism of rule induction is no trivial question and it will need a lot more further research, our entropy model supported by the findings of the current study and our previous findings (Radulescu et al., 2019) allow for the following informed general hypothesis to be formulated about the mechanics and the nature of this *gradual* mechanism. Firstly, if the *input entropy* is low, memorization of the specific items and their probability distribution allows for the input to be encoded by memorization and probability matching to the input. Not only can the specific items and chunks of items be memorized and encoded as per their probability distribution in the input, but the low input entropy allows for *item-bound generalization* as an encoding method, as we have shown in this study. An increase in *input entropy* places a challenge on the (incidental) memorization of the exact items and their surface statistical regularities (i.e. probability distribution), so that the finite rate of information encoding (entropy per second) increases towards its maximum. The higher the individual incidental memorization capacity, the more *input entropy* can be encoded until the finite *channel capacity* is reached. At this point, since the *channel capacity* cannot be exceeded, a change in the encoding mechanism is required in order to enable more input entropy to be encoded, but avoiding exceeding the channel capacity. Thus, the domain-general pattern recognition capacity (a component of the working memory) re-structures the information and groups items into categories in order to reduce the number of bits that each item is coded for, thus compressing the information and encoding it as *category-based generalization.* The better this individual domain-general pattern recognition capacity is, the sooner in the process it can begin to restructure and compress the information.

In conclusion, we suggest that at the computational level (in the sense of Marr, 1982), there is one single mechanism – processing and encoding input entropy by a finite time-dependent entropy processor. However, it is conceivable that, at the algorithmic and implementational levels, different cognitive representations and capacities, and different brain areas take over the details of processing and encoding the bits of information, which needs further research to pinpoint. Thus, while we agree with Marcus and colleagues (1999, 2001, 2012) that the mind shows a symbol-manipulation ability to represent abstract relationships between variables, and to distinguish between mental representations of types and tokens (Marcus, 2001), we think though that this might rather be a mechanistic description of the outcomes of the underlying encoding mechanism.

Furthermore, as we mentioned in the description of our entropy model, sensitivity to entropy entails sensitivity to similarities and differences, which means that our model assumes certain perceptual primitives to be available, though not sufficient, for rule induction, and possibly innate (Endress & Bonatti, 2007; Endress et al., 2009; Marcus, 2001). Moreover, in terms of a biologically plausible efficiency principle, while it can be envisaged that nature endowed the human species with multiple specialized mechanisms (statistical learning for some learning tasks and an innate abstract mechanism that is triggered only when computationally more efficient), we deem a single time-dependent entropy processing mechanism with different outcomes more efficient and plausible. This view is compatible with recent evidence from neurobiology, which converge on the hypothesis that depending on the amount of particular events/data stored or forgotten, the memory system either creates representations that are highly specific to past data – overfitted models – or the memory transience allows for storing less specific past data for the purpose of driving generalization to new and noisy environments (Frankland, Köhler, & Josselyn, 2013; Hardt, Nader, & Wang, 2013; Migues et al., 2016; Richards & Frankland, 2017). Corroborating evidence from neural networks research (Hawkins, 2004; Kumaran et al., 2016; MacKay, 2003) converges on a very similar view: the memory system (and the neural networks as a model) is not only designed for remembering specific data, but also for optimized generalization, by having the capacity to encode a finite degree of specificity or prior knowledge (i.e. entropy, in information-theoretic terms), in order to prevent overfitting to past data for the purpose of allowing for flexibility in noisy environments.

Chapter 4

Size Does Not Matter. Entropy Drives Rule Induction in Non-Adjacent Dependency Learning

Radulescu, S. and Grama, I.[20]

Abstract

In this study, we examined adults' ability to detect and generalize non-adjacent dependencies in an *aXb* grammar under different *input entropy* conditions. We further extend and test an information-theoretic entropy model for rule induction that we proposed in Radulescu et al. (2019). Specifically, our entropy model hypothesizes that an increase in *input entropy* per unit of time gradually adds up to the maximum rate of information encoding (bits/second), i.e. the finite *channel capacity* of the learning system, and causes a change in encoding method in order to avoid exceeding the *channel capacity*. Thus, in this study, we give an extended and more refined information-theoretic approach to a previous *variability hypothesis* that suggested a high number of middle elements (the *X* elements positioned in the middle of the *aXb* strings) is crucial to non-adjacent dependency learning (Gómez, 2002), by showing that it is not the mere set size of the items which drives rule induction in non-adjacent dependency learning, but it is a particular pattern of input variability, i.e. *input entropy*. To this end, we kept the set size constant, and we varied the *input entropy* by manipulating the probability distribution of the items. More precisely, since a large set size of intervening *Xs* was deemed to be a crucial factor in non-adjacent dependency learning, we kept a relatively large set (18 *Xs*) and varied the combinatorial possibilities with three $a_i\_b_i$ frames, so that we obtained three different entropy versions of an a_iXb_i grammar. We found that although the set size of the intervening *Xs* was equally large in all entropy conditions, participants successfully learned the non-adjacent dependencies and generalized them to novel instances better in the highest entropy condition than in the medium and low entropy conditions. Moreover, we found a U-shape pattern of non-adjacent

[20] This chapter is a modified version of a manuscript under review:
Radulescu, S. & Grama, I. (2021) Size Does Not Matter. Entropy Drives Rule Induction in Non-Adjacent Dependency Learning

dependency learning as a function of increasing *input entropy,* with no evidence of learning in the medium entropy condition, consistent with previous findings (Onnis et al., 2003; 2004).

1. Introduction

Non-adjacent dependencies are formally defined as $a_i\_b_i$ frames consisting of frozen cooccurrences between specific *a* and *b* items (or words), which are generalizable over a richer intervening category of *X* elements (or words), such that a_iXb_i triplets are well-formed, while a_iXb_j (where $i\neq j$) are ill-formed For example, in natural languages, such non-adjacent dependencies are deemed to model the mechanism needed for learners to acquire rules like *is* go-*ing, is* learn-*ing,* where *be* always predicts *-ing* over a richer intervening category of verbs.

Previous research has investigated the learning mechanism that supports non-adjacent dependency learning (Frost & Monaghan, 2016; Gómez, 2002; Grama, Kerkhoff, & Wijnen, 2016; Newport & Aslin, 2004; Peña, Bonatti, Nespor, & Mehler, 2002; Romberg & Saffran, 2013; Pacton & Perruchet, 2008; Wang, Zevin & Mintz, 2016; 2019), and proposed several factors to be relevant both to learning the specific non-adjacencies in the input and to generalizing them to novel examples (Wilson et al. (2020), for an extensive review). Peña et al. (2002) suggested that brief pauses are necessary to mark the beginning and the end of the dependencies, as in "word boundaries", in order for learning to be successful, while Endress, Nespor and Mehler (2009) proposed that the non-adjacent *a* and *b* elements must be at the edge of the specific sequence for successful learning. Thus, while one of the early proposed factors was the role of shorter within-word pauses (100–200ms) and longer between-word pauses (800ms) that would signal the a_iXb_i triplets in the input stream (Gómez, 2002; Gómez & Maye, 2005; Gómez, Bootzin, & Nadel, 2006; Romberg & Saffran, 2013), however, more recent findings show that such pauses may not be necessary, since learners were also able to learn the non-adjacencies from a continuous stream (Onnis, Monaghan, Christiansen, & Chater, 2004; Frost & Monaghan, 2016; Wang et al., 2016; 2019).

Another proposed factor was the effect of adjacent dependencies, in that "weaker" adjacent (*aX* and *Xb*) probabilities point the learner towards the "stronger" non-adjacent dependencies, thus prompting learning of specific $a_i\_b_i$ frames over a high-variability intervening *X* category: Gómez (2002) showed that learning was successful only when the set size of the intervening *Xs* was relatively large (i.e. 24 *Xs*), but not when the set size was 2, and the results were inconclusive for set sizes of 6 and 12. These findings led to a prominent *variability hypothesis* on learning of non-adjacencies (Gómez, 2002; Gómez & Maye, 2005): a large set of *Xs,* which renders low transitional probabilities between adjacent elements (*aX* and *Xb*), highlights the $a_i\_b_i$ frames as very predictable dependencies (i.e. higher non-adjacent transitional probabilities between a_i and b_i) which facilitates learning of the specific $a_i\_b_i$ frames. Thus, the increased variability of the intervening *X* elements, which was quantified in terms of the size of the set of specific *X* items*,* was interpreted as the driving factor to render adjacent probabilities unpredictable such that learners would disregard the middle items and their attention would be steered towards the more predictable non-adjacent dependencies. Although, crucially only a *critical mass* of middle elements – 24, not 6 or 12 – i.e. a critical amount of variability was deemed to be a driving factor.

However, more recent findings challenge this account as well, showing that, adjacent and non-adjacent dependencies do not need to compete for learner's attention, and they can be learned simultaneously (Romberg & Saffran, 2013) and, under specific conditions, learning of the non-adjacencies occurs even with a small set size of the intervening *X* elements – 9 *Xs* (Wang et al., 2019) and even 3 *Xs* (Frost & Monaghan, 2016; Wang et al., 2019). Mostly, in these studies learning occurred with a small set size under conditions of continuous speech stream, where learners were simultaneously solving the segmentation task and the non-adjacent dependency learning task. Nonetheless, in a systematic attempt to further investigate and specify the *variability hypothesis,* Onnis and colleagues found evidence for an interesting U-shape pattern of non-adjacent dependency learning as a function of increasing variability of the *X*-set size; in that robust learning of the non-adjacencies was found either under conditions of null variability (i.e. one intervening *X* combined with three $a_i\_b_i$ frames), or under a considerably larger set size of 24 *Xs* (Onnis, Christiansen, Chater, & Gómez, 2003, Onnis et al., 2004). This pattern of results was found both for detecting non-adjacencies in the exposure language, that is learners exposed to a_iXb_i triplets reject a_iXb_j triplets as ill-formed (Onnis et al., 2003), and also for generalizing them to novel sequences, i.e. learners exposed to a_iXb_i triplets reject a_iXb_j triplets as ill-formed, but also generalize $a_i\_b_i$ dependencies to novel a_iNb_i triplets, where *N* stands for new middle items never heard in the familiarization (Onnis et al., 2004).

In order to tease apart the effect of the set size of the intervening *Xs* from the effect of adjacent dependencies (claimed by Gómez, 2002), Wang et al. (2019) used a constant set size of intervening *Xs,* i.e. 9 words, but varied the way the $a_i\_b_i$ frames combined with them, such that in one condition each of the three the $a_i\_b_i$ frames combined restrictively with only a limited set of 3 *Xs*, while in the other condition all the three $a_i\_b_i$ frames combined exhaustively with all 9 *Xs.* Hence, two conditions were created: in the Categorical Condition, learners were exposed to 3 $a_i\_b_i$ * 3 Xs – 9 different triplets, while in the Distributed Condition they listened to 3 $a_i\_b_i$ * 9 Xs – 27 different triplets. As per the authors, this manipulation holds set size of intervening *Xs* constant (9), while the adjacent transitional probabilities were higher in the Categorical Condition (1/3 = 0.33), than in the Distributed Condition (1/9 = 0.11). In the test phase, learners were exposed to a_iNb_i strings, with new middle words unheard in the familiarization, in order to test the hypothesis that the knowledge about non-adjacencies is acquired by generalization of the $a_i\_b_i$ frames over a category of middle elements, rather than by chunk-memorization of the specific triplets. If the theory that low transitional probabilities between adjacent elements facilitate non-adjacent dependency learning (Gómez, 2002) holds true, the Distributed Condition should yield better learning than the Categorical Condition. The results showed learning in both conditions, but no difference between the conditions. The authors concluded that while learners indeed generalized the $a_i\_b_i$ frames to novel sequences in both conditions, there was no evidence for the low adjacent transition probabilities being a facilitating factor for non-adjacent dependency learning. However, the authors did not clearly specify what was in fact the driving factor for non-adjacent dependency learning, since it is not set size of the intermediate *Xs* and not "weaker" transitional probabilities.

In any case, since the set size used by Wang et al. (2019) is rather small – 9 *Xs* – one might argue that a large set size of intervening *Xs* might not be necessary for non-adjacent dependency learning, but it might actually help, as it was the case in Gómez (2002) and in the studies by Onnis and colleagues (Onnis et al., 2003; Onnis et al., 2004)

Another important research question regarding learning non-adjacencies in previous research was whether learners' representations in such studies are actually chunk-like representations, i.e. a_iXb_i sequences of three words memorized as a chunk (Christiansen & Arnon, 2017), or the learning mechanism actually involves generalization of the fixed $a_i\_b_i$ frames over an intervening *X* category. In the latter case, generalization over a category of intervening items means that learners familiarized to a_iXb_i sequences would also accept as grammatical a_iNb_i strings, where the $a_i\_b_i$ frames are generalized over a new intervening word (*N*), which was not heard in the familiarization. Indeed, several studies have shown that learning non-adjacencies does not rely on a chunk-based memorization of the familiarization stimuli, but it involves generalization to novel sequences, i.e. a_iNb_i strings (Frost & Monaghan, 2016; Grama et al., 2016; Wang et al., 2019).

In this article, we further investigate the topic of generalization in non-adjacent dependency learning together with the effect of input variability on non-adjacent dependency learning, by going into a deeper theoretical understanding of the mechanism of generalization and of the particular pattern of variability. In Radulescu, Wijnen and Avrutin (2019), we proposed a general information-theoretic model for rule induction (generalization) in order to investigate the underlying mechanism and factors that drive both *item-bound generalization* and *category-based generalization*, and we applied it to a repetition-based XXY type of grammar (e.g. strings like *da_da_li*). *Item-bound generalization* describes relations that repeatedly occur between specific physical items, i.e. *li* always follows *da* and *da* is always repeated in a sequence, while *category-based generalization* is an operation beyond specific items which describes relations that involve categories (variables), e.g. *Y* always follows *X,* or *X* always follows *da* (where *X* and *Y* are categories taking different several values, and *da* is one specific item). These qualitatively different types of generalization (dubbed in accord with previous suggestions – Gómez & Gerken, 2000) had previously been proposed to reflect two different types of learning mechanisms, with statistical learning underlying *item-bound generalization*, while a higher-order abstract learning mechanism being responsible for the more abstract *category-based learning* (Marcus et al., 1999). However, in accord with a more recent *single-mechanism hypothesis* (Aslin & Newport, 2012; 2014; Frost &

Monaghan, 2016), in Radulescu et al. (2019) we proposed that a single mechanism drives both types of generalization, as a result of a very particular interaction between two factors: the statistical properties of the input, i.e. *input entropy,* and the brain's sensitivity and finite capacity to encode the entropy in the environment, i.e. *channel capacity*. More precisely, Radulescu et al. (2019) proposed that an increase in *input entropy* which is higher than the available *channel capacity* drives the tendency to move from a high-specificity *item-bound generalization* (i.e., in this case, a *same-same-different* rule only with familiar syllables) to a more abstract *category-based generalization* (i.e. a *same-same-different* rule with novel syllables as well). Indeed, Radulescu et al. (2019) exposed adults to a 3-syllable XXY grammar (e.g. strings like *da_da_li, mu_mu_sa*) in six experimental conditions with increasing *input entropy*, and found that adults' tendency to move from an *item-bound generalization* to a *category-based generalization* increased gradually as a function of increasing entropy.

In this article, we further extend our entropy model for rule induction from a repetition-based XXY grammar (Radulescu et al., 2019) to a more complex a_iXb_i grammar, where specific items *a* always predict specific items *b* to create frozen $a_i\_b_i$ frames over a richer intervening category of *Xs.* We suggest that learning of such a complex type of grammar entails both *item-bound generalization* (the dependency between specific *a* and *b* elements), and *category-based generalization* (generalizing the specific $a_i\_b_i$ frames over the intervening category of *Xs*). According to our model, a lower *input entropy* allows for *item-bound generalizations*, while a higher *input entropy* than the finite *channel capacity* drives *category-based generalization*.

Moreover, another goal of this study is to further test the feasibility of entropy as a quantitative measure of input variability. In our previous experiments (Radulescu et al., 2019) we used Shannon's entropy formula (Shannon, 1948), which is a particular function between the number of items and their probability distribution. We created six different entropy versions of the XXY grammar by increasing the number of items, but crucially keeping their probability distribution homogeneous. It might be argued that such a manipulation relies mostly on the increased set size of the items, and less on their probability distribution or the particular relation between the number and the probabilities described by entropy. Therefore, here we manipulate entropy in the opposite way, namely, we keep the set size constant, and we vary the probability distribution. More precisely, since a large set size of intervening *Xs* was deemed to be a crucial factor in non-adjacent dependency learning, we kept a relatively large set (18 *Xs*) and varied the combinatorial possibilities with a classical number of three $a_i\_b_i$ frames, in order to obtain three different entropy versions of an a_iXb_i grammar. We found that although the set size of the intervening *Xs* was constantly large (18) in all conditions, participants successfully learned the non-adjacent dependencies and generalized them to novel instances better in the highest entropy condition than in the medium and low entropy conditions. Thus, in the following section of the paper we elaborate on our entropy model and we formulate specific hypotheses for non-adjacent dependency learning. Next, we present a non-adjacent dependency experiment in which we varied *input entropy* in three conditions, in order to test the specific hypothesis made by our entropy model, and to disentangle the effect of a large set size from the effect of *input entropy* on non-adjacent dependencies. Finally, we conclude the study with the discussion and conclusions sections, where we compare our results with similar results from previous studies, in order to propose a unified account for the underlying mechanism and factors that drive learning and generalization of non-adjacencies.

2. An entropy model for rule induction in non-adjacent dependency grammars

2.1 Brief introduction of the model and previous findings

In Radulescu et al. (2019), we proposed an entropy model which hypothesizes that rule induction is driven by the brain's sensitivity to *input entropy* and its finite encoding capacity, i.e. *channel capacity*. In short (and simplifying for now), less *input entropy* facilitates detecting regularities between specific items, i.e. *item-bound generalization*, while an *input entropy* which is higher than the finite *channel capacity* drives the tendency towards *category-based generalization*. Thus, in accord with the *single-mechanism hypothesis* (Aslin & Newport, 2012; 2014; Frost & Monaghan, 2016), the main tenet of our entropy model is that *item-bound* and *category-based generalizations* are outcomes of the same encoding mechanism, as a reflection of the dynamics between the statistical properties of the input, *input entropy,* and our finite encoding capacity, i.e. *channel capacity*. We define our encoding capacity as *channel capacity,* in information-theoretic terms, which means the finite rate of information encoding (entropy per unit of time), which might be supported by certain cognitive capacities, e.g. memory capacity, in psychological terms.

Taking a step further from other studies that looked into generalization by using similar entropy measures (Ferdinand, 2015; Ferdinand, Kirby, & Smith, 2019; Perfors, 2012; Perfors, 2016; Saldana, Smith, Kirby, & Culbertson, 2017; Samara, Smith, Brown, and Wonnacott, 2017), in Radulescu et al. (2019) and in this study we propose an information-theoretic model that captures the dynamics of the interaction between the *input entropy* and the relevant encoding capacity (i.e. *channel capacity*). Inspired by Shannon's entropy and noisy-channel coding theory (Shannon, 1948), this model specifies a quantitative measure for the likelihood of moving away from encoding specific probability distributions of items to forming more abstract general representations. More precisely, our model hypothesizes that a *bit by bit* increase in *input entropy* per unit of time gradually adds up to the maximum rate of information encoding (bits/second), i.e. the finite *channel capacity* of the learning system. According to Shannon's noisy-channel coding theory, in a communication system, a message (information) is transmitted reliably (that is with the least loss of information to the receiver), if and only if it is encoded by using an encoding method which is efficient enough to keep the rate of information transmission (including the inevitable noise) below the *channel capacity*. Since the finite *channel capacity* acts as an upper bound on the *input entropy* which can be encoded per unit of time, if the input entropy is higher than the channel capacity and as such the message cannot be transmitted reliably using the current encoding method, the need for another more efficient encoding method is created. Thus, it follows that an increase in *input entropy* which is higher than the *channel capacity* should drive the need for another more efficient encoding method. Based on this theory, our entropy model hypothesizes that an increase in *input entropy* renders the encoding method inefficient (that is creating high uncertainty when receiving the message), and drives the transition from *item-bound generalization* to *category-based generalization.*

In Radulescu et al. (2019), we found that increasing the *input entropy* gradually in six experimental conditions (i.e. from 2.8, 3.5, 4, 4.2, 4.58, to 4.8 bits), drives a gradual tendency to move from *item-bound* to *category-based generalization* in an XXY grammar (Radulescu et al., 2019). Learning this type of XXY grammar involves abstracting away from specific items, that is from a *same-same-different* rule with specific syllables occurring in the *X* and *Y* slots (i.e. *item-bound generalization*), and moving to a *category-based generalization,* that is a *same-same-different* rule between the *X* and *Y* categories, regardless of specific items included in these categories. More specifically, successful generalization involves being familiarized with a particular set of 3-syllable XXY strings, e.g. *da_da_li*, and accepting XXY strings with completely novel syllables, which were not present in the familiarization, e.g. *ba_ba_gu.*

2.2 Predictions of the entropy model for non-adjacent dependency learning

While in Radulescu et al. (2019) we probed the effect of *input entropy* on rule induction in a 3-syllable *XXY* grammar, in this study we further develop and test the model by probing the effect of the input entropy on rule induction in a more complex *aXb* grammar. This type of grammar poses a challenge in that successful learners of this type of *aXb* grammar abstract away from an *item-bound* to a *category-based generalization* for the intervening *X* category (Frost & Monaghan, 2016; Grama et al., 2016; Wang et al., 2019), while, crucially, sticking to an *item-bound generalization* for the specific *a_b* dependencies. It can be argued that, while high *input entropy* drives *category-based generalization* for the *X* category, it might impede *item-bound generalization* for the specific *a_b* dependencies of such an *aXb* grammar. Therefore we hypothesize, that the effect of increasing entropy on learning this type of grammar is not a *gradually* better performance as we found for an XXY grammar (Radulescu et al., 2019), but there might be a particular (critical) amount of *input entropy*, that is a lower and an upper bound on the *input entropy* (which we hypothesize is determined by the *channel capacity*), such that an interaction between *input entropy* and *channel capacity* facilitates detection of the specific *a_b* dependencies and generalizing them over the category of intervening *Xs.*

More specifically, here are the main hypotheses of our entropy model, as we stated them generally in Radulescu et al. (2019), to which we add here more specifications relevant for non-adjacent dependency learning:

1. **Lower *input entropy*** than the *channel capacity* facilitates encoding the information by a method which matches the probability distribution of the specific items. Thus, if the *input entropy* is lower than the *channel capacity,* the information about specific items and their configuration (i.e. entropy of the input) can be reliably transmitted through the channel (i.e. with the least loss of information at receiver's end) at the available channel capacity (i.e. the maximum rate of information encoding), and encoded by *item-bound generalization.* This hypothesis predicts that in the case of non-adjacent dependencies, a low *input entropy* allows for specific relations between *a* and *b* elements to be readily detected and encoded by matching their configuration. That means the exact $a_i\_b_i$ frames can be detected and encoded by *item-bound generalization,* i.e. specific *a* items (a_i, a_j) *always* pair with specific *b* items (b_i, b_j).

2. **Higher *input entropy*** than the *channel capacity* drives a change in the encoding method, such that the information can be reliably transmitted through the channel at the available *channel capacity*. This hypothesis is based on the noisy-channel coding theory (Shannon, 1948), according to which, if the *input entropy* is higher than the *channel capacity*, another more efficient encoding method can be found, but the rate of transmission (*input entropy per second*) cannot exceed the *channel capacity.* Thus, based on these concepts, if the *input entropy* increases, the *item-bound generalization* becomes inefficient and prone to errors (i.e. causes high loss of information) due to the upper bound placed by the *channel capacity*, which cannot be exceeded. As a consequence, another more efficient encoding method needs to be found, in order to avoid exceeding the *channel capacity,* which would cause great loss of information. As we argued in Radulescu et al. (2019), it is this essential feature of the *channel capacity* which precipitates restructuring of the information, such that the item-specific features and their configuration are (unconsciously) reobserved by identifying similarities/differences in order to compress the message by a more efficient encoding method. As a result, insignificant differences between items (i.e. specific low-probability features) are erased or "forgotten", which in turn flashes out non-specific shared features between items and facilitates grouping them in categories based on these shared features (Radulescu et al., 2019). This hypothesis predicts that in the case of non-adjacent dependencies, high *input entropy* drives restructuring of the information and shapes *item-bound generalization* into *category-based generalization,* such that the intervening rich set of *Xs* is encoded as a category, and the $a_i\_b_i$ frames can be generalized over the intervening *X* category. Fast and reliable (i.e. with least loss of information) encoding of the intervening elements as a category *X*, that is as a compressed message which reduces the amount of bits/s needed to encode the intervening elements, provides enough capacity (i.e. *channel capacity* in bits/s) to encode the $a_i\_b_i$ frames reliably as *item-bound generalizations.*

3. Since our entropy model predicts a gradual transition from *item-bound* to *category-based generalization*, as a function of increasing *input entropy,*

medium entropy creates an environment which does not facilitate *item-bound generalization,* thus there is high uncertainty due to loss of information when encoding specific items and their configuration. Also, medium entropy is not high enough to drive *category-based generalization,* such that category formation is also incomplete and creates uncertainty. In Radulescu et al. (2019) we argued and showed that a medium entropy environment, where none of the two encoding methods is strongly developed, i.e. not highly efficient at encoding the input, will result in the most uncertain situation for the learners, creating thus an overall drop in the learning curve. Thus, also in the case of an *aXb* grammar, which requires both *item-bound* and *category-based generalization*, we expect a medium entropy environment to create the most uncertain situation, because *item-bound generalization* would be too weak to clearly highlight mismatches between the specific a_i and b_i items, and *category-based generalization* is not fully developed to facilitate category formation of the intervening *Xs.* Thus, we expect a drop in performance in medium entropy, compared to the high entropy environment, since medium entropy is not high enough to drive *category-based generalization*, but it is high enough to interfere with *item-bound generalization.*

This prediction was borne out in the medium entropy conditions of Radulescu et al. (2019), where we found a moderate tendency towards *category-based generalization*, that is to generalize the *same-same-different* rule to novel XXY strings with unfamiliar syllables. We also found a U-shape pattern of results for the Familiar-syllable X1X2Y test items as a function of increasing *input entropy* (Radulescu et al., 2019): more explicitly, successful correct rejection of the Familiar-syllable X1X2Y test items was supported either by a strongly developed *item-bound generalization* in low entropy conditions, which highlights mismatches with the *same-same-different* rule with familiar syllables, or by a strongly developed *category-based generalization* in high entropy conditions, which highlights violation of the *same-same-different* rule regardless of familiar or new syllables. However, under medium entropy conditions, where none of the two encodings was fully developed, we found the lowest tendency to correctly reject the Familiar-syllable X1X2Y strings as ungrammatical. Hence, we found a U-shape pattern of results as a function of increasing *input entropy* for the testing condition that is theoretically supported by a strong development of either one or the other encoding method. Based on those findings and the hypotheses of our entropy model, we predict that a medium entropy environment for non-adjacent dependency learning in an *aXb* grammar would create an uncertain situation which would lead to inefficiency of both encoding methods, such that confident detection and generalization of the non-adjacency would be impeded.

Our entropy model takes a step further from previous theories that claimed the set size of the intervening *Xs* plays a crucial effect on non-adjacent dependency learning (Gómez, 2002; Gómez & Maye, 2005), such that a large set size of intervening *Xs* lowers the adjacent transitional probabilities and highlights the more predictable dependencies between the less variable *a* and *b* elements. Intuitively, a large set size creates more entropy, since entropy is a function between the number of items and their probability distribution. In

Gómez (2002), the set size of the intervening *Xs* was also paired with a uniform probability distribution of the items, such that all *a_b* frames were exhaustively combined with all the *Xs* rendering a uniform probability distribution of the *a_b* frames and *Xs*. Thus, we suspect that the actual factor that drove better learning in the large set size condition in Gómez (2002) was *input entropy*, as hypothesized by our model, but it was not immediately evident, since learning could be interpreted as an effect of set size of intervening *Xs*. Our entropy model makes the prediction that the opposite manipulation of entropy, i.e. a uniformly large set size, but with a skewed probability distribution (not uniform probability distribution) lowers the entropy. If indeed the factor at stake is *input entropy,* we expect that high *input entropy* drives better learning than lower entropy, in spite of a uniformly large set size. Thus, in this study we aimed to tease apart the effect of set size of the intervening *Xs* from the effect of *input entropy*, by keeping a large set size of *Xs* (18) constant and varying the probability distribution to obtain three different entropy conditions.

3. Experiment: design and rationale

The goal of this study is two-fold:

1. On the one hand we aimed to assess the generalizability of our model to more complex non-repetition grammars. To this end, we employed an *aXb* grammar, which is based on learning the specific $a_i\_b_i$ frames (i.e. *item-bound generalization*), and generalizing them over a richer intervening category of *X* elements (i.e. *category-based generalization*).

2. The second goal was to further address and investigate the suitability and feasibility of entropy as a quantitative measure of input complexity, which has an effect of rule induction. In our previous study (Radulescu et al., 2019), we used Shannon's entropy, in order to capture a specific pattern of variability, which entails a particular relation between the number of items in a set and their probability distribution. In our previous study, we created six different entropic levels by increasing the number of items of each set, but keeping a uniform probability distribution. Readers might argue that such a manipulation of entropy relies mostly on the mere number of items in the set, and less on the probability distribution or the particular relation between number and probability distribution described by the entropy formula. Thus, in this study, we created three entropy versions of the aXb grammar by manipulating entropy in the opposite way: we kept the number of items constant, and we varied their probability distribution.

Firstly, to address our first goal, we aimed at further investigating the *variability hypothesis*, by seeking a better understanding of the effect of variability on rule induction in non-adjacent dependencies, specifically by using *input entropy* as a measure of input variability. To the best of our knowledge, this is the first study that specifically applies and tests entropy-based hypotheses on rule induction in non-adjacent dependencies. Onnis et al. (2003) argued for a very specific effect of variability on non-adjacent dependency learning, namely that either null variability or very high variability of the intervening *X* elements would facilitate detection of non-adjacencies, while medium variability would lead to the worst performance (i.e. the authors argue in favor of a U-shape pattern as a function of increasing variability). This prediction was based on the rationale that learners are actively (though, unconsciously) looking for sources of reduction in uncertainty, such that they would entertain the most predictable and invariant structure in the input: when the large intervening *X*-set is very unpredictable, the more predictable $a_i\_b_i$ frames "flash out" and can be easily learned. Or, conversely, the authors argue that null variability in the middle elements make the a_iXb_i frames stand out because of their variability. In any case, small -to medium-sized sets of *Xs* (2 – 6) might confuse learners because both the middle elements and the outer frames vary, but none of them significantly more than the other (Onnis et al., 2003).

However intuitive and logical this account in favor of a U-shape pattern might sound, nevertheless, there is a theoretical weakness that we would like to help strengthen, and also an experimental weakness in the findings reported by the authors, that we would like to also further address. Firstly, the account seems to inconsistently suggest that learners are either (unconsciously) looking towards a source of less variability (more predictable $a_i\_b_i$ frames compared to middle *Xs*) when *X*-set is large, or towards a source of more variability (more variable $a_i\_b_i$ frames compared to the single *X*) when *X*-set is only one. This would actually suggest an inconsistent behavior, if we see unpredictability in the *X*-set we look for predictability in the $a_i\_b_i$ frames, but if we find predictability in the *X*-set we look for unpredictability in the $a_i\_b_i$ frames. Secondly, robust learning under null variability was found for detecting non-adjacencies in the familiarization stimuli, but not for generalizing them to novel items, that is learners exposed to a_iXb_i triplets reject a_iXb_j triplets as ill-formed, exclusively with familiar *Xs* (Onnis et al., 2003), which is highly likely to have been supported by plain rote memorization of the 3 a_iXb_i triplets repeated 144 times (Experiment 1, in Onnis et al., 2003) and not by actual learning of a non-adjacency rule (i.e. *item-bound generalization* in our terminology). In their subsequent study, Onnis et al. (2004) address this confound by also adding a test for generalization of the non-adjacencies to novel sequences, i.e. to test whether learners exposed to a_iXb_i triplets also generalize $a_i\_b_i$ dependencies to novel a_iNb_i triplets. However, we think that this test was not particularly convincing since 9 out of 12 test items could have been correctly answered based only on a mismatch with the memorized a_iXb_i triplets (the test consisted of 3 a_iXb_i triplets, 3 a_iXb_j triplets, 3 a_iNb_i triplets, and 3 a_iNb_j triplets), such that the observed learning effect might be mostly carried by those 75% of (mis)matched strings (3 a_iXb_i triplets, 3 a_iXb_j triplets and 3 a_iNb_j triplets) against the memorized ones from familiarization (i.e. 3 a_iXb_i triplets). Hence, we think further evidence is needed in order to clearly investigate the *null variability effect* of the intervening *Xs.*

Therefore, in this study we aim at further extending and fine-tuning the *variability hypothesis* by offering a more refined information-theoretic model which gives a consistent theoretical account for the *variability effect* and the previously observed U-shape pattern of performance.

In order to address the second goal of this study, we used monosyllabic Dutch-like nonsense words for the *a* and *b* elements, and bisyllabic Dutch-like nonsense words for the intervening *X* elements. The *aXb* grammar generated three entropy versions by splicing the syllables into *aXb* strings according to the following combinatorial rules. Unlike Gómez (2002), we kept *X* set size constant (18 *Xs*) and varied entropy by combining each of the three *a_b* frames with different (sub)sets of *Xs*, as follows. In order to obtain the low entropy version, each of the three *a_b* frames was combined with a distinct subset of 6 *Xs* (3 *a_b* * 6 *Xs*), such that the subsets of *Xs* did not overlap, which generated a rather low entropy version (H_L = 3.52 bits). For the medium entropy, the *aXb* grammar combined the three *a_b* frames with partially overlapping subsets of 12 *Xs* (3 *a_b* * 12 *Xs*), which generated a medium entropy version (H_M = 4.27 bits). In order to generate the high entropy version, the *aXb* grammar combined exhaustively each of the three *a_b* frames with the entire set of intervening *Xs* (3 *a_b* * 18 *Xs*), which yielded a rather high entropy (H_H = 4.7 bits). Since such evaluations of low/medium/high entropy could be deemed as relative, depending on the grammar/language, we based our estimates of the set size and variability of such an *aXb* grammar on previous studies on non-adjacent dependency learning (Gómez, 2002; Grama et al., 2016). For the entropy calculations, we employed the same implementation model as in Radulescu et al. (2019) – see Table 1 below for complete entropy calculations.

High Entropy	Medium Entropy	Low Entropy

H[begin-*a*]=H[3] = -Σ[0.333*log0.333] = 1.58
H[aX] = H[54] = 5.75
H[Xb] = H[54] = 5.75
H[*b*-end] = H[3] = 1.58
H[begin-aX] = H[54] = 5.75
H[aXb] = H[Xb-end] = H[54] = 5.75
H[bigram] = 3.67
H[trigram] = 5.75
H[total] = $\frac{\text{H[bigram]+H[trigram]}}{2}$ = **4.71**
	H[begin-*a*]=H[3] = -Σ[0.333*log0.333] = 1.58
	H[aX] = H[36] = 5.17
	H[Xb] = H[36] = 5.17
	H[*b*-end] = H[3] = 1.58
	H[begin-aX] = H[36] = 5.17
	H[aXb] = H[Xb-end] = H[36] = 5.17
	H[bigram] = 3.36
	H[trigram] = 5.17
	H[total] = $\frac{\text{H[bigram]+H[trigram]}}{2}$ = **4.27**
		H[begin-*a*]=H[3] = -Σ[0.333*log0.333] = 1.58
		H[aX] = H[18] = 4.17
		H[Xb] = H[18] = 4.17
		H[*b*-end] = H[3] = 1.58
		H[begin-aX] = H[18] = 4.17
		H[aXb] = H[Xb-end] = H[18] = 4.17
		H[bigram] = 2.86
		H[trigram] = 4.17
		H[total] = $\frac{\text{H[bigram]+H[trigram]}}{2}$ = **3.52**

Table 1. Entropy values for the three entropy versions of the *aXb* grammar
(TP = 1/6 = 0.16 in Low Entropy, TP = 1/12 =0.083 in Medium Entropy, TP = 1/18 =0.055 in High Entropy)

In the test phase, participants were asked to provide grammaticality judgements on *aXb* strings with either correct (familiar) or incorrect (unfamiliar) *a_b* frames. Importantly, all (correct and incorrect) test strings included new middle *X* elements. If learners correctly accept *aXb* strings with the correct *a_b* frames and new *X* elements, it shows they were both able to encode *item-bound generalizations* (i.e. the *a_b* frames) and to generalize them over a category of *X* elements, i.e. *category-based generalization*.

4. Methods

4.1 Participants

76 healthy, non-dyslexic Dutch speaking participants (53 females, age 18 – 51, M = 22) were assigned to either the High, Medium or Low condition. Four more participants were tested in the High Entropy condition, but subsequently excluded due to self-reported prior familiarity with non-adjacent dependency learning experiments. Only healthy participants that had no known language, reading or hearing impairment or attention deficit were included. They all signed a consent form and were paid for their participation.

4.2 Materials

Familiarization stimuli. All *a* and *b* elements were monosyllabic nonsense words resembling Dutch phonotactics (e.g., *tɛp, jɪk*), while all *X* elements were bisyllabic Dutch-like nonsense words (e.g., *naspu, dyfo:*). All *a* and *b* elements were recorded in *aXb* strings, read out in a lively, child-friendly intonation. The *X* elements were recorded as direct object nouns in Dutch carrier sentences ("Ik zie de _ in de tuin", "I see the _ in the garden"). This resulted in *aXb* strings where the *a* and *b* elements were very salient in terms of pitch (see Grama, Kerkhof & Wijnen, 2016, Emphasized 250ms Condition).

In order to obtain three entropy versions with low, medium and high entropy, the *aXb* grammar spliced each *a_b* frame with a (sub)set of the *X* elements (see Appendix A for the complete list of the familiarization X elements), such that the set of *Xs* was constant (18 *Xs*) in all versions, but entropy varied as follows. For the low entropy, each of the three *a_b* frames was combined with one subset of 6 *Xs* (3 *a_b* * 6 *Xs*), and each of these strings was repeated 18 times (i.e. 324 strings in total). For the medium entropy version, the *aXb* grammar combined the three *a_b* frames with partially overlapping subsets of 12 *Xs* (3 *a_b* * 12 *Xs*), and each of the strings was repeated 9 times to yield the same total number of strings. In order to generate the high entropy version, the *aXb* grammar combined exhaustively each of the three *a_b* frames with the entire set of intervening *Xs* (3 *a_b* * 18 *Xs*), with 6 repetitions of each string.

In all entropy conditions, we used two versions of the *aXb* language: Language 1 (L1) and Language 2 (L2). The only difference between L1 and L2 was the specific legit combination of the three *a* and *b* elements into frames: in L1 the grammatical frames were *tɛp _lyt, sɔt_ jɪk* and *rak_tuf*, while in L2 the grammatical frames were *tεp _ jɪk, sɔt_tuf* and *rak_lyt*. Therefore, every a_i _b_i pair in L1 was ill-formed (a_i_b_j) in L2 and vice versa. The reason for using these two different versions of the *aXb* grammar was to exclude a possible effect of idiosyncrasies of particular *a_b* combinations (L1 or L2) on learning. Each participant was randomly assigned to only one version of the *aXb* grammar (either L1 or L2), and randomly assigned to only one entropy condition (either Low or Medium or High Entropy), such that we obtained a between-subjects experimental design. We employed a within-string pause of 250ms between the elements of a string (*a, X, b*) as well as a between-string pause of 750ms.

Test stimuli. For the test stimuli, the *aXb* grammar spliced each *a_b* frame of each language (L1, L2) with two novel *X* elements to yield (6 *a_b* * 2 *X* =) 12 new test items (see Appendix A for the test *X* elements). Thus, the six new *aXb* strings with the *a_b* frames of L1 were ungrammatical for the participants exposed to L2, while the six new *aXb* strings with the *a_b* frames of L2 were ungrammatical for the participants exposed to L1. As such, each participant was exposed to 12 new *aXb* strings, six of which were grammatical and six ungrammatical. Accuracy score for the learning of the *aXb* grammar was calculated as the mean correct answers, i.e. correct acceptance of the grammatical test strings and correct rejection of the ungrammatical test strings.

4.3 Procedure

Participants were seated in a sound-proof booth. Before the familiarization phase they were instructed that they would listen to an "alien language" that does not resemble any language that they might be familiar with, and which has its own words and grammar. To avoid any motivation to explicitly look for patterns in the stimuli, participants were not informed of the subsequent test phase until after the end of the familiarization phase. Participants were also given the simultaneous task of coloring a mandala while listening to the "alien language", in order to promote implicit learning and to prevent explicit attention directed to the structure of the input. Before the test phase, participants were instructed that they would listen to new sentences in the same "alien language", none of which would be identical to the sentences they had heard before. They were then asked to decide for each sentence whether it was correct or not, according to the grammar of the language they had just heard, by clicking on "Yes" or "No". They were instructed to answer quickly and intuitively. The experiment lasted around 15 minutes.

4.4 Results

Figure 1 presents the mean accuracy, that is percentage of acceptances of correct *aXb* strings and rejections of incorrect *aXb* strings per group, across the three conditions (Low Entropy, Medium Entropy and High Entropy). The mean accuracy in the High Entropy condition was *M* = .65 (*SD* = .24), in the Medium

Entropy condition it was *M* = .49 (*SD* = .13), while in the Low Entropy condition it was *M* = .54 (*SD* = .14).

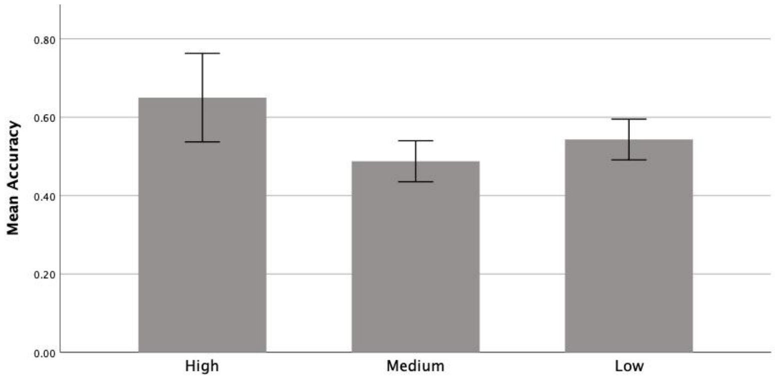


Figure 1. Proportions of correct acceptance for grammatical items and correct rejection of ungrammatical items, in each Entropy condition. Error bars show 95% Confidence Intervals of the mean.

Error Bars: 95% CI

Figure 2 shows the distribution of individual mean rates in each experimental condition, High Entropy, Medium Entropy and Low Entropy.

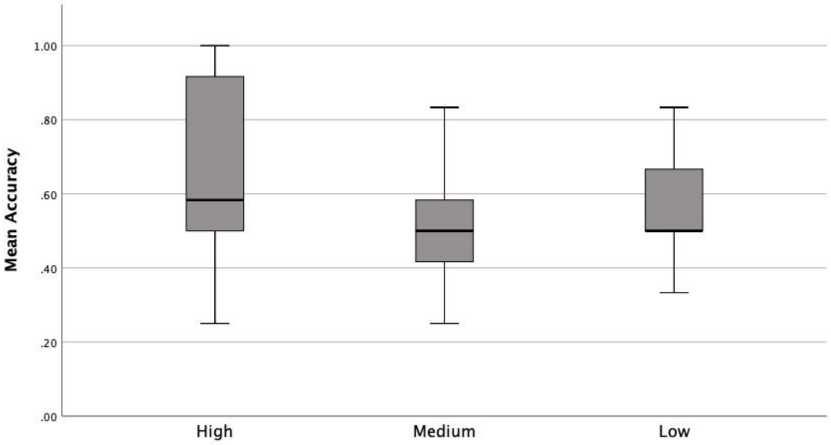


Figure 2. The distribution of individual mean accuracy rates per test type in each condition: High, Meadium and Low Entropy

In the High Entropy Condition a One-Sample Kolmogorov-Smirnov Test showed that the accuracy rates were not normally distributed (*p* = .022), so we ran a One-Sample Wilcoxon Signed-Rank Test, which showed significant above-chance performance (Median = 58%, SE = 15.839, Z = 2.336, *p* =. 019). In the Medium Entropy Condition a One-Sample Kolmogorov-Smirnov Test showed normally distributed data (*p* = .07), so we ran a One-Sample T-Test, which showed that accuracy rates were not significantly different from chance (Mean 49%, SD =

13.21, *p* = .631). In the Low Entropy Condition a One-Sample Kolmogorov-Smirnov Test revealed that data was not normally distributed (*p* <.001), so we ran a One-Sample Wilcoxon Signed-Rank Test which showed a marginally-significant above-chance performance (Median =.50%, SE = 22.553, Z = 1.709, *p* = .088).

In order to probe the effect of *input entropy* on rule induction, we compared the performance in the three conditions (High, Medium and Low Entropy groups) in a general linear mixed effects analysis (using IBM SPSS version 26) of the relationship between Accuracy (correct acceptance of the grammatical test items and correct rejection of the ungrammatical ones) and Entropy Condition (High, Medium and Low Entropy groups). Therefore, as dependent (binomial) variable we entered Accuracy score into the model. As fixed effects we entered Entropy Condition (High, Medium and Low Entropy), Language (L1, L2) and an Entropy Condition x Language interaction. As random effect we had an intercept for subjects. Because Language (L1, L2) did not improve the model and it did not show a significant effect or interaction with Entropy Condition, we excluded it from the final analysis. 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. The model reported here is the best fitting model, both in terms of the model's accuracy in predicting the observed data, and in terms of AIC (Akaike Information Criterion).

We found a significant main effect of Entropy Condition (*F* (2, 909) = 5.441, *p* = .004), and planned Bonferroni comparisons yielded a significant difference between High and Medium Entropy (*t* (1) = 3.279, *p* = .001, 95% CI [0.279 ; 1.111]), a significant difference between High and Low Entropy (*t* (1) = 2.261, *p* = .024, 95% CI [0.062 ; 0.883]), and no significant difference between Medium and Low Entropy (*p* = .238).

Moreover, performance across the different entropy conditions resulted in a U-shaped function, with a polynomial trend analysis showing a significant quadratic effect (F(2, 73) = 5.366, *p* = .007.

Further, Cohen's effect size value for the mean difference in correct answers between the High and Medium Entropy was d = .83 (large effect size) with an effect size correlation *r* = .38, between the High and Low Entropy it was d = .56 (medium effect size), *r* = .27, while between the Low and Medium Entropy it was d = .37 (small effect size), *r* = .18.

5. Discussion and Conclusions

In this study, we probed adults' ability to learn non-adjacent dependencies when exposed to an a_iXb_i grammar under different *input entropy* conditions, in order to assess the generalizability of our entropy model to more complex non-repetition grammars (Radulescu et al., 2019; 2020). More precisely, we aimed at giving a more fine-tuned and refined information-theoretic approach to a previous prominent hypothesis that suggested a certain pattern of input variability to be a driving factor in learning non-adjacencies (Gómez, 2002; Gómez & Maye, 2005). Previously it was suggested that the set size of the a_iXb_i

grammar, particularly a large set size of the intervening *X* elements facilitates learning of the $a_i\_b_i$ dependencies, due to low adjacent transitional probabilities which highlight and shift attention to the "stronger" dependencies between the non-adjacent *a* and *b* elements.

To this end, we manipulated the *input entropy* of a non-adjacent dependency a_iXb_i grammar by keeping the set size equally large, but varying the probability distribution of the items. More specifically, since a large set size of intervening *Xs* was deemed to be a crucial factor in non-adjacent dependency learning (Gómez, 2002; Gómez & Maye, 2005), we kept a relatively large set (18 *Xs*) and varied the combinatorial possibilities with three $a_i\_b_i$ frames. We obtained three different entropy versions of an a_iXb_i grammar: a low entropy version, a medium entropy version and a high entropy version. We found that although the set size of the intervening *Xs* was equally large in all entropy conditions, participants successfully detected the non-adjacent dependencies and generalized them to novel instances better in the highest entropy condition than in the medium and low entropy conditions. We interpret these findings to be in line with the hypothesis of our entropy model, namely that high *input entropy* drives restructuring of the information and, thus, reliable encoding (i.e. with least loss of information) of the middle elements as a category *X*. This *category-based generalization* is in fact a compressed form of encoding, which reduces the amount of bits/s needed to encode the intervening elements, thus providing enough remaining capacity (i.e. *channel capacity* in bits/s) to encode the $a_i\_b_i$ frames reliably as *item-bound generalizations.*

Next, according to our entropy model, an *input entropy* which is below the *channel capacity* allows for encoding the information by matching the probability distribution of the specific items in the input, namely the specific $a_i\_b_i$ frames can be detected and encoded by *item-bound generalization,* i.e. specific *a* items (a_i, a_j) *always* pair with specific *b* items (b_i, b_j). Although we did not find robust learning in the low entropy condition, we think this was the case because although we dubbed it as the lowest entropy condition compared to the other two entropy versions we investigated, it actually falls in a rather medium entropy range (3.52 bits).

Moreover, in the medium entropy condition we found no evidence of learning the non-adjacencies, although the set size was equally large in all the entropy conditions. These findings bring further evidence in favor of the hypothesis of our entropy model which predicts that a medium entropy environment, where none of the two encoding methods is highly efficient at encoding the input (causing high loss of information), results in the most uncertain situation for the learners (Radulescu et al., 2019). Similarly, in the case of an *aXb* grammar, which requires both *item-bound* and *category-based generalization*, we interpret our findings of the lowest performance in the medium entropy condition to be in line with our hypothesis: a medium entropy environment causes a drop in the learning curve, because *item-bound generalization* is too weak to clearly highlight mismatches between the specific a_i and b_i items, and *category-based generalization* is not fully developed to facilitate category formation of the intervening *Xs.* Further evidence for this hypothesis was brought by our finding of a U-shape pattern of results as a function of increasing *input entropy.* Further research should look into the U-shape curve of learning, in order to confirm these results, if better accuracy would be found with a much lower input entropy.

To summarize, all these findings are in line with the main hypotheses of our entropy model regarding rule induction in a non-adjacent dependency grammar. Specifically, these results bring strong evidence to the hypothesis that it is not the mere set size of the intervening *X* elements, but rather variations in *input entropy* that drive rule induction in non-adjacent dependency learning. Also, these results bring further evidence in favor of the previously found U-shaped curve of learning (Onnis et al., 2003, 2004, 2015). The main contribution of this study is showing that the U-shape pattern of learning is not an effect of mere set size of items, but an effect of the *input entropy.*

Another goal of this study was to manipulate entropy in a different way from the one we employed in Radulescu et al. (2019), in order to further probe the feasibility of entropy as a measure of input variability. Specifically, instead of keeping a uniform probability distribution of items and increasing their number, in the present study we did the opposite: we kept a constantly large number of the intervening *Xs* and we varied the probability distribution, in order to obtain different levels of *input entropy*. Results showed that even when having a large set size, if entropy is not high enough, rule induction is impaired. Thus, we interpret these results to show that indeed entropy is a suitable quantifying method for assessing the effect of input variability on rule induction, and thus, *input entropy*, not the mere set size, is the driving factor in rule induction in non-adjacent dependency learning.

Our findings add to and take a step further from another study that aimed at giving a more in-depth analysis of the input variability as a factor in learning non-adjacencies: Wang et al. (2019) found evidence for non-adjacent dependency learning even under conditions of small set size (9 *Xs*). However, they did not find better learning when adjacent transitional probabilities were lower rather than higher, which was previously claimed to be the case (Gómez, 2002; Gómez & Maye, 2005). In any case, since the set size used by Wang et al. (2019) is rather small – 9 *Xs* – one possible hypothesis could be that a large set size of intervening *Xs* might not be necessary for non-adjacent dependency learning, but it might boost non-adjacent dependency learning, as it was the case in Gómez (2002), Onnis et al. (2003) and Onnis et al. (2004). Thus, we investigated that possibility in this study by employing a larger set (18 *Xs*) and found *input entropy* to be a driving factor rather than large set size. We suggest an alternative interpretation to Wang et al.'s findings, according to our entropy model: even though adjacent transitional probabilities were lower in the Distributed Condition, the increase in *input entropy* was not sufficient to drive a significantly higher tendency to generalize the $a_i\_b_i$ frames to novel sequences in the Distributed Condition (4.35 bits) as compared to the Categorical Condition (3.17 bits – see Appendix B for the entropy calculations for the experimental condition in Wang et al. (2019) according to the implementation method we proposed in Radulescu et al. (2019)). This finding is consistent with our findings reported in this article: learning in the medium entropy condition (4.27 bits) was not better than it was in the low entropy condition (3.52 bits), but a further increase in *input entropy* up to 4.71 bits lead to significantly better learning than in both medium and low entropy conditions. In any case, based on these findings, we suggest that the effect of *input entropy* on rule induction in non-adjacent dependency might not be linear or gradual, but in stages, precisely due to the nature of such a complex a_iXb_i grammar, which requires a certain balance between *item-bound generalization* and *category-based generalization*. Further research is needed in order to specify more precisely this hypothesis, and the amount of predicted increase in *input entropy* which is necessary for significantly better learning.

When comparing the results we obtained in the low entropy condition of this study with the results of Radulescu, Kotsolakou, Wijnen, Avrutin and Grama (2021), where we employed a low entropy version of the same grammar with the same *input entropy* as the low entropy version used in this study – 3.52 bits, we find an interesting pattern of results. While in this study we did not find robust learning of the non-adjacent dependencies in the 3.52-bit version (M = .54, SD = .14, *p* = .088), in the Radulescu et al. (2021) we did find significant learning in the 3.52-bit version (M = .69, SD = .46, *p* = .001). The two experiments had the same design, stimuli and combinatorics to obtain the same input entropy, but the only crucial difference was the number of repetitions of each *aXb* string: in the present study, each of the three *a_b* frames was combined with one subset of 6 *Xs* (3 *a_b* * 6 *Xs*) to obtain 18 different strings (*X* set size = 18), and each of these strings was repeated 18 times (i.e. 324 strings in total), while in Radulescu et al. (2021), each of the three *a_b* frames was also combined with one subset of 6 *Xs* (3 *a_b* * 6 *Xs*) to obtain 18 different strings (*X* set size = 18), but crucially they were repeated only 12 times, resulting in a total of 216 strings. In information-theoretic terms, this makes a difference in the inflow of information per unit of time (bits/s), which may or not reach the maximum rate of information transmission (i.e. *channel capacity*).

Specifically, we interpret the difference in these results to point to the essential role played by the *channel capacity* in rule induction. In the present study the inflow of information was 3.52 bits/symbol in 15 minutes, while in Radulescu et al. (2021) the inflow of information was 3.52 bits/symbol in 10 minutes. Thus, in the present study, one minute of familiarization with this grammar version carries on average 3.52/15 = 0.234 bits/symbol of information, while in Radulescu et al. (2021), one minute of familiarization with the same grammar version carries on average 3.52/10 = 0.352 bits/symbol of information. In simple words, in this study, the information is more diluted, more dispersed over symbols in time, which makes it more likely to remain below the available *channel capacity.* On the other hand, in Radulescu et al. (2021), the information is more compressed, more concentrated in symbols over time, which makes it more likely to be higher than the available *channel capacity,* and as a result to drive better generalization of the non-adjacent dependencies. This is, of course, only a possible logical explanation, but which deserves more future research in order to further specify the hypothesis of an interaction between *input entropy* and our time-dependent *channel capacity* in non-adjacent dependency learning.

Appendix A

	a/b	***IPA***
a1	tep	[tɛp]
a2	sot	[sɔt]
a3	rak	[rɑk]
b1	lut	[lyt]
b2	jik	[jik]
b3	toef	[tuf]
	X	
No.	***Familiarization***	***IPA***
1	blieker	[blikər]
2	dufo	[dyfo]
3	fidang	[fidɑŋ]
4	gopem	[xopəm]
5	kengel	[kɛŋəl]
6	kijbog	[kɛibɔx]
7	loga	[loxa]
8	malon	[malɔn]
9	movig	[movix]
10	naspu	[nɑspu]
11	nijfoe	[nɛifu]
12	noeba	[nuba]
13	plizet	[plizɛt]
14	rajee	[raje]
15	rogges	[rɔxəs]
16	seeta	[seta]
17	snigger	[snixər]
18	wabo	[vɑbo]

	Test (novel Xs)	
19	nilbo	[nilbo]
20	pergon	[perxɔn]

Appendix B

Categorical Condition	**Distributed Condition**

H[***b***-*a*]=H[9] = 3.17
H[aX] = H[9] = 3.17
H[Xb] = H[9] = 3.17
H[*b*-***a***] = H[9] = 3.17
H[begin-aX] = H[9] = 3.17
H[aXb] = H[Xb-end] = H[9] = 3.17
H[bigram] = 3.17
H[trigram] = 3.17
H[total] = $\frac{\text{H[bigram]+H[trigram]}}{2}$ = **3.17**
	H[***b***-*a*]=H[9] = 3.17
	H[aX] = H[27] = 4.75
	H[Xb] = H[27] = 4.75
	H[*b*-***a***] = H[9] = 3.17
	H[begin-aX] = H[27] = 4.75
	H[aXb] = H[Xb-end] = H[27] = 4.75
	H[bigram] = 3.96
	H[trigram] = 4.75
	H[total] = $\frac{\text{H[bigram]+H[trigram]}}{2}$ = **4.35**

Entropy calculations for the two conditions of the *aXb* grammar employed by Wang et al. (2019) in Experiment 4. Since the *aXb* strings were played in a continuous stream, i.e. without pauses between triplets, there are no cues as to the beginning and ending of the triplets, thus the first bigram consists of the *b* element of the previous triplet and the *a* element of the current triple, and the last bigram consists of the *b* element of the current bigram and the *a* element (similar to the calculations for transitional probabilities in continuous streams).

Chapter 5

Fast But Not Furious. When Sped Up Bit Rate of Information Drives Rule Induction

Radulescu, S., Kotsolakou, A., Wijnen, F., Avrutin, S. and Grama, I.[21]

Abstract

Young and adult learners' abilities range from memorizing specific items to finding statistical regularities between them (*item-bound generalization*) and abstracting away from the input to apply general rules to novel instances (*category-based generalization*). Both external factors, like input variability, and internal factors, like cognitive limitations, have been shown to be driving factors of learners' ability to form general representations from exposure to a limited set of examples. Yet the exact dynamics between these factors and the circumstances under which rule induction emerges remain largely underspecified. In this article we further extend our information-theoretic model (Radulescu et al., 2019) – based on Shannon's noisy-channel coding theory (Shannon, 1948) – which adds into the "formula" for rule induction the crucial dimension of *time* and rate of information transmission, i.e. the rate of encoding information by a time-sensitive encoding mechanism. Specifically, our model hypothesizes that, if the *input entropy per second* is higher than the maximum rate of information transmission (bits/second), which is determined by the *channel capacity,* the encoding method moves gradually from *item-bound generalization* to a more efficient *category-based generalization*, so as to avoid exceeding the *channel capacity*. Thus, the goal of this study is two-fold. The first goal is theoretical, since to the best of our knowledge, this is the first study that specifically tests a hypothesis based on Shannon's *channel capacity* and the *noisy-channel coding theory* in artificial grammar learning. To this end, we first define and give a concrete example/proposal of how *channel capacity* can be estimated in an artificial grammar learning experiment (our experiments from Radulescu et al., 2019). More precisely, we show evidence that the rate of information transmission reached the *channel capacity* in Radulescu et al.'s study (2019) and that, as predicted by our model, the transition from one encoding method to another more efficient encoding method (*category-based generalization*) is signaled by an initial increase followed by a decrease in rate of

[21] This chapter is a longer version of a manuscript under review:
Radulescu, S., Kotsolakou, A., Wijnen, F., Avrutin, S. & Grama, I. (2021) Fast But Not Furious. When Sped Up Bit Rate of Information Drives Rule Induction

equivocation (i.e. loss of information), calculated from learners' performance. The second goal of this study is to take the next logical step of directly manipulating the *time* variable of the *channel capacity*. To this end, in two artificial grammar experiments with adults we sped up the bit rate of information transmission, crucially not by simply reducing the time between stimuli by an arbitrary amount, but by a factor that we calculated based on data from our previous experiments, by using the *channel capacity* formula. We found that when we increased the bit rate of information transmission in a repetition-based XXY grammar, learners' tendency towards *category-based generalization* increased, as predicted by our model. Conversely, we found that increased bit rate of information transmission in a more complex non-adjacent dependency *aXb* grammar led to poorer learning in general, at least judging by our specific way of assessing accuracy. This finding could be accounted by our model, since speeding up the bit rate of the inflow of information precipitates a change from *item-bound generalization* into a *category-based generalization,* which means that it impedes *item-bound generalization* of the specific *a_b* frames, but it facilitates *category-based generalization* for the intervening *X* category of elements, and possibly categories of *a* and *b* elements, instead of specific dependencies between specific *a* and *b* elements.

1. Introduction

An increasing body of evidence has showed that both young and adult learners possess a domain-general distributional learning mechanism that enables them to find statistical patterns in the input (Saffran, Aslin, & Newport, 1996; Thiessen & Saffran, 2007), and also a learning mechanism that allows for category (rule) learning (Marcus et al, 1999; Smith & Wonnacott, 2010; Wonnacott, 2011; Wonnacott & Newport, 2005). Rule induction (generalization or regularization) has been often explained as resulting from processing the variability in the input (quantifiable amount of statistical variation), both in young and adult language learners (Gerken, 2006; Hudson Kam & Newport, 2009; Hudson Kam & Chang, 2009; Reeder, Newport, & Aslin, 2013).

This study looks into the factors that drive the inductive step from memorizing items and statistical regularities to inferring abstract rules. While previously cognitive psychology theories claimed that there are two qualitatively different mechanisms, with rule learning relying on encoding linguistic items as abstract categories (Marcus et al, 1999), as opposed to learning statistical regularities between specific items (Safran et al., 1996), recent views converge on the hypothesis that one mechanism – *statistical learning* – underlies both item-bound learning and rule induction (Aslin & Newport, 2012; 2014; Frost & Monaghan, 2016; Radulescu, Wijnen & Avrutin, 2019). Aslin & Newport (2014), in particular, argue that the evidence from relevant studies (Gerken, 2006; Reeder et al., 2013) supports a *single-mechanism hypothesis* with the finding of *a gradient of generalization*. According to the authors of these studies, learners show a different pattern of learning, depending on the input variability: they either generalize ("abstract rule learning") or they withhold generalization, such that it apparently leads to only "surface statistical learning" (Aslin & Newport; 2012; 2014).

While supporting the *single-mechanism hypothesis* proposed previously, in Radulescu et al. (2019) we took a step further in understanding the underlying mechanism and the *gradient of generalization*. This study investigates the two qualitatively different forms of learning discussed by previous research mentioned above, which were dubbed, in accord with previous suggestions (Gómez and Gerken, 2000), *item-bound generalizations* and *category-based generalizations.* Specifically, Radulescu et al. (2019) suggest that, unlike in previous studies, the underlying processes should be disentangled from their outcomes, that is the learning mechanisms (*statistical learning* and *abstract rule learning)* should be conceptualized separately from the resulting forms of encoding (*item-bound generalizations* and *category-based generalizations*).

While *item-bound generalizations* describe relations between specific physical items, e.g. a relation based on physical identity, like "*ba* follows *ba*", *category-based generalizations* are operations beyond specific items that describe relations between categories (variables), e.g. "Y follows X", where Y and X are variables taking different values. In order to explain *how* and *why* a single mechanism outputs these two qualitatively different forms of encoding, Radulescu et al. (2019) proposed an information-theoretic model of rule induction as an encoding mechanism. In this model, we put together both the statistical properties of the input, i.e. *input entropy*, and also the brain's capacity to encode the input under conditions of finite cognitive capacities. We define our encoding capacity as *channel capacity,* in information-theoretic terms, which means the finite rate of information encoding (entropy per unit of time), which might be supported by certain cognitive capacities, e.g. memory capacity, in psychological terms.

Indeed, previous research hinted at potential cognitive constraints, i.e. memory limitations, on rule learning: the *Less-is-More* hypothesis (Newport, 1990; 2016) proposed that the differences in tendency to generalize between young and adult learners stem from the maturational differences in memory development. Specifically, limited memory capacity leads to difficulties in storing and retrieving low-frequency items, which prompts overuse of more frequent forms leading to overgeneralization. Consequently, a few studies investigated the exact nature of these cognitive constraints and the exact mechanisms by which they operate and affect rule induction. They showed that, while there is some evidence for the *Less-is-More* hypothesis on memory constraints modulating rule induction (Hudson Kam & Chang, 2009; Hudson Kam & Newport, 2005; Hudson Kam & Newport, 2009; Wonnacott, 2011), it is not yet clear under *what* specific circumstances and *why* memory constraints should have a certain effect on rule learning (Perfors, 2012). The *Less-is-More hypothesis* was also investigated in terms of its domain-generality, and studies found that cognitive constraints are also reflected in the regularization behavior in non-linguistic domains (Kareev, Lieberman, and Lev, 1997; Ferdinand, Kirby, and Smith, 2018), while other studies found that the regularization tendencies and patterns are very similar across domains and language levels – morphology vs word order (Saldana, Smith, Kirby, & Culbertson, 2017).

Nevertheless, the exact cognitive load and mechanisms at stake in rule induction have yet to be thoroughly specified and understood. To this end, Radulescu et al. (2019) offer an extended and more refined information-theoretic approach to the *Less-is-More hypothesis*, by proposing an entropy model for rule induction, which quantifies the specific pattern of statistical variability in the input (i.e. *input entropy* – measured in bits of information) to which the brain is sensitive. Our model hypothesizes that rule induction is driven by the interaction between the input entropy and the finite encoding capacity of the brain (i.e. *channel capacity*). Crucially, the model proposes that rule induction is an automatic process that moves *gradually* – *bit by bit* – from a high-fidelity item-specific encoding (*item-bound generalization*) to a more general abstract encoding (*category-based generalization*), as a result of the input entropy being higher than the *channel capacity*, i.e. the maximum rate of information encoding (bits/s).

The model is based on Shannon's *entropy* and *channel capacity* concepts (Shannon, 1948), and Shannon's noisy-channel coding theory, which says, in short, that in a communication system, a message (or information) can be transmitted reliably (i.e. with the least loss of information), if an only if encoded by using an encoding method that is efficient enough so that the rate of information transmission (i.e. per unit of time), including noise, is below the channel's capacity. If the rate of information transmission (bit rate) is higher than the *channel capacity*, then another more efficient encoding method can be found, but the *channel capacity* cannot be exceeded.

Based on these concepts, our entropy model for rule induction posits that the change in the encoding method – from *item-bound* to *category-based generalization* – is driven by a kind of a regulatory mechanism, which moves from an inefficient encoding method (i.e. with high loss of information), to a more efficient encoding method, which allows for higher input entropy to be encoded reliably (with the least loss of information possible) per unit of time, but crucially below the channel's capacity. Loss of information (or uncertainty at learner's end, in information-theoretic terms) should be understood as the missing bits of information caused by the noise interference during transmission through the channel in time. Thus, this model adds into the rule induction "formula" the crucial dimension of *time*, i.e. the rate of encoding information by a time-sensitive encoding mechanism, and consequently the decrease of system's loss of information in time by moving to a more efficient encoding.

Indeed, previously a few studies used different (not information-theoretic) ways of quantifying and manipulating a time-dependent variable, in order to investigate the role it plays in category learning (longer exposure time – Endress & Bonatti, 2007; Reeder, Newport, & Aslin, 2009; 2013), in non-adjacent dependency learning (faster speech rate – Wang, Zevin & Mintz, 2016; 2019) and in auditory statistical learning (temporal distance between successive stimuli – Emberson, Conway & Christiansen, 2011). Even though all these studies used different designs, different stimulus material, and different approaches to the temporal variable they manipulated, nevertheless, a clear pattern stands out, namely that generally a shorter time is beneficial to auditory rule (category) learning. However, the exact amount of time, the mechanism and the reasons for it having a positive effect on rule learning are still to be fully investigated and understood.

In order to answer these questions, this study further extends the entropy model we proposed in Radulescu et al. (2019), and puts forth an innovative information-theoretic approach to the time-dependent variable, that is not by an arbitrary manipulation of the amount of time between stimuli or exposure time, but by employing the information-theoretic concept of *channel capacity* and Shannon's noisy-channel coding theory.

To sum up, the goal of this study is two-fold: theoretical, since to the best of our knowledge this is the first study that applies Shannon's concept of *channel capacity* and his noisy-channel coding theory to artificial grammar learning. Specifically, we will further extend our entropy model by defining *channel capacity,* and giving a concrete example of how we can apply and estimate Shannon's *channel capacity* in artificial grammar learning (e.g. our experiments from Radulescu et al., 2019). The second goal of our study is experimental, namely we will take the next logical step of directly manipulating the time-dependent variable of the *channel capacity* in two other artificial grammar experiments, which specifically test *channel capacity* hypotheses posed by our entropy model.

Therefore, we will first briefly introduce our entropy model supported by our previous findings, then we will present the *channel capacity* and Shannon's noisy-channel coding theory. Next, we will give an example and a brief proof of concept, by showing how *channel capacity* and the rate of information transmission can be applied and quantified in an artificial language learning environment for rule induction, i.e. our previous experiments reported in Radulescu et al. (2019). These quantifications enable an estimation of *channel capacity,* which we will use in the second part of the current study in order to manipulate experimentally an increase in source rate of entropy per second that we can safely assume to be higher than the estimated rate of the *channel capacity.* Specifically, we will present two artificial grammar experiments, in which we speed up the bit rate of information transmission by a factor that we calculated from data obtained in our previous experiments, by using the information-theoretic concepts of *channel capacity* and rate of information transmission.

2. An Entropy Model for Rule Induction

Among other studies that used entropy measures to look into regularization patterns (Ferdinand, 2015; Ferdinand, Kirby, & Smith, 2019; Perfors, 2012; Perfors, 2016; Saldana, Smith, Kirby, & Culbertson, 2017; Samara, Smith, Brown, and Wonnacott, 2017), Radulescu et al. (2019) and the present study take a step further and propose an information-theoretic model that captures the dynamics of the interaction between the *input entropy* and the encoding capacity (*channel capacity*). This model specifies a quantitative measure for the likelihood of transitioning from encoding specific probability distributions to category formation. Specifically, our model hypothesizes that the *gradient of generalization* (Aslin & Newport, 2012) results from a *bit by bit* increase in *input entropy* per unit of time, which gradually adds up to the maximum rate of information transmission (bits/s), i.e. *channel capacity* of the learning system.

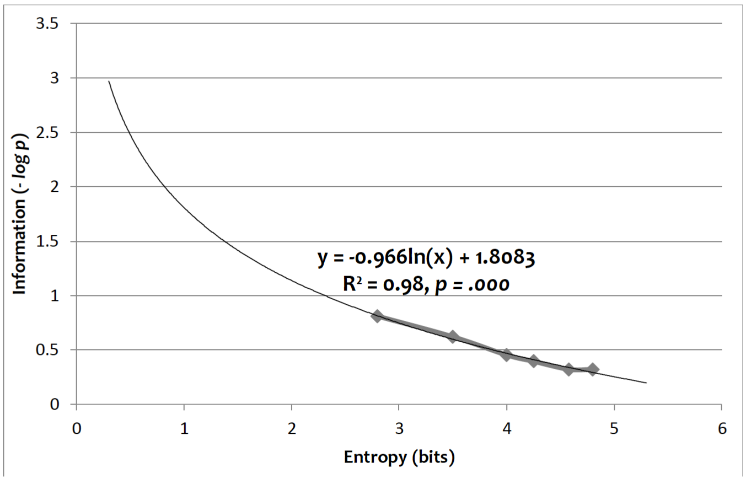
Given a random variable *X*, with *n* values {$x_1, x_2 \ldots x_n$}, Shannon's entropy (Shannon, 1948), denoted by *H(X)*, is defined as:

$H(X) = -\sum_{i=1}^{n} p(x_i) log p(x_i)$ [22];

where *p(x$_i$)* is the occurrence probability of x_i. This quantity (H) measures the information per symbol produced by a source of input, i.e. it is a measure of the average uncertainty (or surprise) carried by a symbol produced by a source, relative to all the possible symbols (values) contained by the set (Shannon, 1948).

In Radulescu et al. (2019), in two artificial grammar experiments, we exposed adults to a 3-syllable XXY artificial grammar. We designed six experimental conditions with increasing input entropy (2.8, 3.5, 4, 4.2, 4.58, 4.8 bits). Results showed that an increase in input entropy *gradually* shaped *item-bound generalization* into *category-based generalization* (Radulescu et al., 2019). Thus, we obtained a precise measure of learner's sensitivity to the input entropy: learner's information load (=surprise) about the XXY structure decreases logarithmically as the input entropy increases (Fig. 1). These findings bring strong evidence for the *gradient of generalization* depending on the probabilistic properties of the input, as proposed by Aslin & Newport (2014).

While in Radulescu et al. (2019) we probed the effect of the first factor (*input entropy*), in this study we further develop and test the model by probing the effect of the second factor – *channel capacity* – on rule induction.



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

Figure 1. Information load of the learner regarding *XXY* rule per input entropy. Taken from (Radulescu et al., 2019)

2.2 Channel capacity in information-theoretic terms

This section elaborates on the other factor of our entropy model, namely *channel capacity,* which is another information-theoretic concept closely related to entropy in Shannon's *noisy channel coding theory* of a communication system. Shannon (1948) defines a communication system as having five main components: an information source (which produces a message), a transmitter (which encodes the message into a signal), a channel (the medium used to transmit the signal), a receiver (which does the inverse operation of the transmitter, that is decodes the signal to reconstruct the message), and a destination (the person or thing for which the message is intended). In simple words, an information source produces a message, which is encoded by a transmitter into a signal to be transmitted to a destination. The information transmission occurs via a medium, i.e. a channel of transmission, and it reaches a receiver, which performs the decoding operation on the signal in order to reconstruct the message to deliver to the destination. The main factor under investigation here is the medium used for the transmission of the information, i.e. the *channel,* and its capacity for information transmission. It follows, and it must be specified that the process of information transmission encompasses all processes starting with the information transmission from the source to the destination, that is all the transmission and encoding – decoding processes.

Now, having briefly described the process of information transmission, we can move on to define the *channel capacity.* In order to do so, we first have to define the two main factors that are relevant for the *channel capacity*: the source rate of information transmission and the noise. Since the process of information transmission occurs in time, Shannon defined the concept of source rate of information transmission, which is the amount of information that a source transmits per unit of time. Since information is measured by using *entropy*, the *source rate of information transmission* (H') is the amount of entropy that the source produces per unit of time (bits/s), or the source rate of information production.

Another important aspect in Shannon's theory was the fact that the ideal case of a noiseless transmission is nearly impossible to achieve under normal real-life conditions. Thus, when defining the communication channel, Shannon also took into account another variable, i.e. the *noise*. In Shannon's communication theory, noise is defined as any random perturbations that interfere with the signal, thus rendering a *noisy channel*. The noise might perturb the signal during transmission through the channel or at either terminal end, i.e. transmitter and receiver's end. As a result, there is *uncertainty* when decoding the sent signal and reconstructing the message, which results from the missing bits of information due to a noisy transmission. Shannon (1948) defined this uncertainty, which is in fact a loss of information, as *rate of equivocation (E)*. The rate of equivocation (i.e. missing bits of information) can be minimized by certain methods, but it cannot be reduced to zero by any operation performed on the received signal, thus it is not possible in general to reconstruct the transmitted signal with total certainty (Shannon, 1948).

After having defined all these concepts, the actual rate of information transmission (R) in a noisy environment can be obtained by subtracting the rate of equivocation (E) from the source rate of information transmission – H' (Shannon, 1948):

R = H' – E.

Note that the *actual rate of information transmission* (R) is different from the *source rate of information transmission* (H'), since it takes into account the information loss due to noise (E), which occurs in the transmission of information from the source to the destination. The *source rate of information transmission* (H') is the rate at which the source produces and transmits information (i.e. the source rate of information production), while the *actual rate of information transmission* (R) is quantified at the other terminal end, i.e. the receiver, after the noise had caused a loss in information (E).

Having described and defined the communication channel and the process of information transmission, we can go on to define the factor at stake in our entropy model, namely *channel capacity* (C). Shannon (1948) argued and demonstrated mathematically that the capacity of a noisy channel should be the maximum possible rate of information transmission (R), which can be obtained if and only if the encoding method is adequate and efficient:

C = Max (R) = Max (H' – E).

In other words, the maximum rate of information transmission, i.e. the *channel capacity,* can be achieved by employing an adequate and efficient method of encoding, such that the rate of equivocation (E) is kept at a minimum, so that the rate of information transmission is as close as possible to the source rate of production. That means that the received signal will be as close as possible to the sent signal, and consequently the message will be received with the least uncertainty (i.e. the least loss of information).

According to Theorem 11 by Shannon (1948), given a certain source with a rate of information production H' (entropy per unit of time), if H' < C, information can be sent through a noisy channel at the rate *C* with an arbitrarily small frequency of errors by using a proper encoding method. If H' > C, it is possible to find an encoding method to transmit the signal over the channel, such that the rate of equivocation is minimum, as specified by Shannon – less than H' – C + *e* (*e stands for errors)*, but the rate of transmission can never exceed C. If there is an attempt to transmit a message at a higher rate than C, by using the same encoding method, then there will be an equivocation rate at least equal to the excess rate of transmission. In other words, this means that a message can only be communicated reliably if it is encoded in such a way, i.e. using an efficient encoding method, so that the rate of information transmission, including noise, is below the capacity of the channel.

It follows that, the efficiency of the encoding method is defined by the ratio of the actual rate of transmission to the capacity of the channel. If the encoding method is maximally efficient, the equivocation rate (E) is minimum, so the actual rate of transmission (R) approaches its maximum, which is the channel capacity: C = Max(H' - E) = Max(R). In the ideal noiseless case (where E = 0), R/C = 1, because R = C. If the encoding method is less than maximally efficient, the equivocation rate is higher than 0 (E > 0), so R is lower than C, thus, R/C < 1. In other words, an encoding method is efficient if the equivocation rate is minimum in order for the rate of transmission to achieve its maximum to match the channel capacity. If the rate of equivocation increases, the rate of transmission decreases, which drives the need for finding a more efficient encoding method, in order to achieve a lower rate of equivocation.

It follows that the birth of a new more efficient encoding method is signaled by an initial increase of the rate of equivocation, followed by a decrease in the rate of equivocation, which shows the system has found an encoding method which allows for the maximum rate of information transmission to be reached. We will come back to show how this prediction can be probed experimentally in an artificial grammar learning environment, in section 2.4, where we give a brief proof of the concept of *channel capacity,* and we show information-theoretic evidence for the transition to a more efficient encoding method, based on our previous experiments from Radulescu et al. (2019).

2.3 Main hypotheses of the model about the effect of channel capacity on rule induction

Before going into more details about *channel capacity* in information-theoretic terms, here are the main general hypotheses of the entropy model about the effect of *channel capacity* on rule induction:

1. *Item-bound generalization* and *category-based generalization* are outcomes of the same information encoding mechanism that *gradually* goes from a high-specificity form of encoding (*item-bound generalization)* to a more general abstract encoding (*category-based generalization)*, as triggered by the interaction between *input entropy* and the finite encoding capacity of the learning system. The encoding mechanism moves from an *item-bound* to a *category-based generalization* as the *input entropy per unit of time* increases and becomes higher than the maximum rate of information transmission, i.e. the *channel capacity,* as follows:

a. If the source rate of information transmission (H' – that is the average entropy produced by the source per second – input entropy per second) is below or matches the *channel capacity*, then the information can be encoded using an encoding method which matches the statistical structure of the input, i.e. the probability distribution of the specific items in the input. Thus, if H'≤C, the information about specific items with their uniquely-identifying (acoustic, phonological, phonotactic, prosodic, distributional, etc.) features and their probability distribution (i.e. input entropy) can be encoded with a high-fidelity item specificity, and transmitted through the channel, with little loss of information, at the channel rate – the maximum rate of information transmission – and encoded by *item-bound generalization.* If H'>C, *item-bound generalization* is impeded.

b. If an attempt is made to exceed the finite *channel capacity* of the encoding system, that is the source rate of information transmission (H' – *input entropy per second*) does not match the *channel capacity*, but it is higher than the *channel capacity*, it is possible to find a proper method that encodes more information (entropy), but the rate of information transmission cannot exceed the available *channel capacity*. According to Theorem 11 (Shannon, 1948), if there is an attempt to transmit information at a higher rate than C, by using the same encoding method, then there will be an equivocation rate at least equal to the excess rate of transmission. In other words, the increased source rate of information (H'>C) brings higher inflow of *noise*, which interferes with the signal and causes an increased equivocation rate or information loss (as explained above). Thus, we hypothesize that it is precisely the *finite channel capacity* which drives restructuring of the information, in order to find another more efficient encoding method. A more efficient encoding allows for higher input entropy per second to be encoded reliably (with the least information loss possible).

As we argued in Radulescu et al. (2019), information is re-structured by (unconsciously) re-observing the item-specific features and the structural properties of the input. Noise introduces random perturbations that interfere with the signal and the feature configuration. This leads to instability, which unbinds features and sets them free to interact and bind into new structures. Thence, similarities (shared features) which have a higher significance (i.e. are 'stronger' due to their higher probability) are kept in the new encoding, while differences between items (unshared features), which are insignificant features (e.g. low probability 'noisy' features) are erased or 'forgotten'. This leads to a compression of the signal by reducing the number of unshared 'noisy' features encoded with individual items (i.e. bits of information) and grouping them in 'buckets' (categories). As a result, a new form of encoding is created, which allows for higher *input entropy* to be encoded using the available *channel capacity*, thus yielding a more general (less specific) *category-based* encoding method*.* Thus, the *finite channel capacity* is designed to drive re-structuring of the information for the purpose of adapting to noisier (=increasingly entropic) environments, by the principle of self-organization in line with Dynamic Systems Theory invoked in studies of other cognitive mechanisms, e.g. Stephen, Dixon, and Isenhower (2009).

2. *Channel capacity* is used here to model the encoding capacity used in linguistic rule induction, in information-theoretic terms (i.e. at the computational level, in the sense of Marr (1982))[23]. In psychological terms (at the algorithmic level), we follow experimental evidence from the *Less-is-More hypothesis* line of research, which suggests that memory constraints drive

[23] Although with different definitions and applications, *channel capacity* has previously been used in early work on capacity in memory studies in psychology (Miller, 1956) and in more recent mathematical modelling for inferring workload capacity using response time hazard functions (Townsend & Ashby, 1978;Townsend & Eidels, 2011).

linguistic rule induction (Hudson Kam & Newport, 2005; Hudson Kam & Newport, 2009), and we embed this in classical and recent models of memory and attention (Baddeley, Eysenck, and Anderson, 2015; Cowan, 2005; Miller, 1956; Oberauer & Hein, 2012). Hence, we hypothesize that the cognitive capacities that underlie *channel capacity,* specifically in linguistic rule induction (and, implicitly, in category formation), are the attentionally-controlled regions of activated long-term memory, in other words working memory (WM). Rule induction can be argued to rely on the storage and online time-dependent processing capacities that support the ability to maintain active goal-relevant information (the rule) while concurrent processing (of other possible hypotheses, and of noise) takes place (which is what defines WM as well – Conway et al., 2002). Corroborating evidence comes from positive correlations found between WM and domain-general categorization tasks (Lewandowsky, 2011).

Thus, while we generally deem linguistic rule induction to be supported by a domain-general WM capacity, rather than language-specific algebraic rule learning as proposed by early prominent research (Marcus et al., 1999), in the current study we are exploring specific possible WM components directly involved in linguistic rule induction, besides more general storage and retrieval components tested in previous studies under the *Less-is-More hypothesis* (Hudson Kam & Chang, 2009; Perfors, 2012). Hence, we specifically predict that one of the components underlying *channel capacity* in linguistic rule induction is a domain-general pattern recognition capacity, given that a rule induction task can be intuitively envisaged as a task of finding patterns/rules in the input.

A possible candidate test of domain-general pattern recognition is the Raven's Standard Progressive Matrices (RAVENS – Raven, Raven, & Court, 2000), which was shown to be based on rule induction (Carpenter, Just & Shell, 1990; Little, Lewandowsky, & Griffiths, 2012) and to rely on similar storage and online time-dependent processing capacities to maintain active goal-relevant information (the rule) while concurrent processing takes place (Conway et al., 2002). Although this pattern recognition test and WM capacity are not identical (Conway et al., 2003), and apparently WM is not a causal factor for pattern recognition either (Burgoyne, Hambrick, & Altmann, 2019), high positive correlations were found between measures of WM capacity and tests for this domain-general pattern-recognition capacity (like RAVENS – e.g. Conway et al., 2002; Little, Lewandowsky and Craig, 2014; Dehn, 2017).

3. A developmental increase of *channel capacity*, (e.g. resulting from growth/development of the underlying cognitive capacities) entails higher amount of entropy that can be encoded per unit of time, and thus it reduces the need and the tendency to move to a higher-order *category-based* form of encoding. Thus, if young and adult learners are exposed to the same input entropy, young learners will have a higher tendency to encode the input as *category-based generalization* than adults, because young learners' *channel* has a lower information encoding rate. There is experimental evidence from the *Less-*

is-More hypothesis line of research in favor of our hypothesis, according to which limited memory capacity in young learners might lead to difficulties in storing and retrieving low-frequency items, therefore prompting overuse of more frequent forms, which leads to overgeneralization (Hudson Kam & Newport, 2005; Hudson Kam & Newport, 2009).

2.4. Channel capacity and rate of transmission in an artificial grammar experiment. A Brief Proof of Concept

After having briefly presented a communication system in Shannon's terms and having defined the key concepts and potential psychological mechanisms implementing key notions, let us next describe the process of artificial grammar learning as such a system, in order to offer a brief proof of concept regarding the effect of *channel capacity* on rule induction. In this section, we first describe the experiments by Radulescu et al. (2019) in Shannon's information-theoretic terms of a communication system, then we show how the rate of equivocation and the maximum rate of information transmission, i.e. *channel capacity*, can be estimated in these experiments. Finally, we show how we can probe experimentally a specific prediction (which we briefly formulated in section 2.2) that follows from the *noisy channel capacity* hypothesis (1.b). This prediction specifies what exactly signals a change into a more efficient encoding method, such that we can conclude that the *item-bound generalization* transitioned to the *category-based generalization,* due to a higher source rate of information transmission than the available *channel capacity.* Namely, in accord with Shannon's definition of *channel capacity* and Theorem 11, we predict that the birth of a new more efficient encoding method (as defined previously in section 2.2) is signaled by an initial increase of the rate of equivocation, followed by a decrease of the rate of equivocation.

Prediction: *when increasing source rate of information transmission, if we obtain an increase followed by a decrease of rate of equivocation, it means there was* ***indeed*** *a change in encoding method which was caused by a* ***higher source rate of information transmission than the available channel capacity****.*

In other words, the initial increase of the rate of equivocation caused by an increase in the source rate of information transmission (H') shows that the old encoding method is no longer efficient for reliable information transmission, (i.e. the loss of information due to noise is very high). The subsequent decrease in the rate of equivocation, shows that in order to cope with a higher input entropy per second than the available *channel capacity*, the system found a new encoding method which allows for the maximum rate of information transmission to be reached. Here we show an innovative way to calculate and measure experimentally the increase and decrease of the rate of equivocation (i.e. the loss in bits of information against the sent signal, which creates learner's uncertainty about the message) in order to estimate the *channel capacity*, and to show (in information-theoretic terms) the transition from *item-bound generalization* to *category-based generalization* in artificial grammar learning.

It is important to disambiguate the meaning of the word *uncertainty*: this should not be understood in psychological terms, i.e. as the condition of being in doubt. The meaning of this term in this chapter, and throughout this dissertation, is the information-theoretic meaning of entropy, as defined by Shannon (1948).

To this end, let us first describe an artificial grammar experiment in Shannon's information-theoretic terms of a communication system, by taking the example of our experiments from Radulescu et al. (2019): an artificial grammar (the source) produces a miniature XXY language (the message). The message is a stream of clusters of acoustic frequency patterns perceived as syllables, with a syllable structure that observes Dutch phonotactics, and the stream is structured consistently in 3-syllable strings with very specific combinatorial properties for the bigrams and trigrams of syllables: [XX] – first bigram has a syllable and its duplicate, [XY] – the second bigram has two different syllables, [XXY] – the trigram has a *same-same-different* structure of syllables.

A pseudo-artificial[24] language system (the transmitter) encodes this message into a signal (signal = the set of all possible XXY strings that could belong to the language). In Radulescu et al. (2019), we used six signal versions (signal version = a particular set out of the possible XXY strings): S = {S1, S2, … S6}. All six signal versions have the *same-same-different* structure, but a different entropy at the level of bigram/trigram combinatorics (H_S = {H1, H2, … H6} = {2.8, 3.5, 4, 4.2, 4.58, 4.8 bits}). In terms of experimental design, each signal version (S = {S1, S2, … S6}) corresponds to one of the six sets of stimuli presented in each experimental condition in Radulescu et al. (2019). The signal entropy increases from S1 to S6, depending on the number of bigrams/trigrams and their probability distribution. More specifically, the probability of each particular bigram/trigram decreases across signal versions, since the set of discrete symbols (i.e. particular bigrams/trigrams) is an increasingly large set. As a result, each particular bigram/trigram becomes less significant in the transmission of the message. Thus, the encoding method becomes increasingly efficient from S1 to S6, as it transmits the message (i.e. XXY language) using an increasing number of discrete symbols, which highlights the fact that the message is an abstract *same-same-different* language, regardless which discrete symbols are employed by the signal.

The signal is sent through the channel (learner's learning system) which can be envisaged as a system of several channels as follows: the acoustic signal made of structured clusters of frequency patterns is transmitted via the perception system (the perception channel) and decoded (by the receiver of the perception system, say the "phonological channel") into a stream of phonemes structured in syllables with particular combinatorial properties (as described

[24] We are dubbing it "pseudo-artificial language system", since it is not purely artificial, as it imitates some properties of natural languages and it was created manually by the authors/experimenters, not by a machine.

above), which is the received signal (with some amount of noise from the transmitter – receiver path). This output signal of the phonological channel becomes the input signal to the cognitive system, where another communication system takes over the signal and the function of information transmission. In this communication system the information source is the phonological system sending a signal to the cognitive system, and so the process repeats. The signal is received by the receiver of the cognitive system, which decodes the signal in order to reconstruct the message, i.e. to create representations.

After having modeled the artificial grammar learning as a communication system in Shannon's information-theoretic terms, we can now turn to estimating the rate of equivocation and the maximum rate of information transmission, i.e. *channel capacity*, in these experiments. By employing Shannon's formula presented in section 2.2 above (C = Max (H' – E)), the rate of transmission of the signal in these experiments can be calculated from the source rate of information transmission (H') and the rate of equivocation (E) associated with the received signal.

Firstly, we show how to calculate the source rate of information transmission (H') for each of the six signal versions. The input entropy was transmitted at a relatively slow rate (compared to natural speech rate), but a commonly used rate in artificial grammar experiments (i.e. 50ms within-string pause and 750ms between-string pause, which yielded a total presentation time of approximately 70s per each exposure phase – there were three exposure phases for each version of the grammar). In information-theoretic terms, the source was transmitting approximatively 0.6 bigrams/s and 0.4 trigrams/s, so on average 0.5 symbols/s, where 'symbol' stands for an abstract unit (variable) of statistical information relevant for the learner in this case (i.e. an average of bigrams/trigrams per second). By multiplying this source rate (0.5 symbols/s) by the entropy per symbol (H), we can calculate the entropy per second which was transmitted in each condition of the experiments, using Shannon's formula for the source rate of information production (H'), i.e. the source bit rate:

$H' = mH$,

where *m* stands for the average rate of symbols sent by the source per second, and H is the entropy per symbol.

Since *m* = 0.5 was the same in all six versions of the signal, for H_S = {H1, H2, … H6} = {2.8, 3.5, 4, 4.2, 4.58, 4.8 bits}, the source rate of information transmission is H'_S = {mH1, mH2, … mH6} = {1.4, 1.75, 2, 2.13, 2.29, 2.4 bits}.

Secondly, the equivocation rate (E) is the uncertainty when receiving the signal and decoding it in order to reconstruct the message. The entropy of the sent signal or input to the channel – H(x) – and the entropy of the received signal – H(y) – are equal only if the transmission through the channel is noiseless. If the channel is noisy, as it is the case in nearly all real-life cases, there is a loss of information so that the entropy of the received signal is not equal to the entropy of the sent signal, which leads to uncertainty when decoding the signal. Shannon (1948) argued and demonstrated mathematically that the only and proper way to quantify this uncertainty is by calculating the conditional entropy of the message – Hy(x) – that is the average ambiguity of the received signal or the equivocation when receiving the signal with H(y), when in fact a signal with H(x) was sent.

Next, given a source with a rate of transmission H', and an equivocation rate of Hy(x) per symbol (or per second, *mHy(x)*), the actual rate of information transmission (R) can be calculated by subtracting the rate of equivocation from the rate of production of the source (Shannon, 1948):

$R = H' - E = H' - mHy(x)$.

In simple words, since the rate of equivocation actually quantifies the missing information in the received signal, the actual rate of information transmission (R) is the entropy per second sent by the source minus the missing bits of information due to a noisy channel.

So the formula for the channel capacity of a noisy channel, which is the maximum possible rate of information transmission, can be re-written as:

$C = \text{Max}\ (H' - mHy(x))$.

The calculations for the source rate of transmission were shown above ($H'_S = \{mH1, mH2, \ldots mH6\}$), and now we make a proposal about how to estimate the rate of equivocation, i.e. the conditional entropy of the received signal by the learners of the XXY language, when each signal version ($S = \{S1, S2, \ldots S6\}$) was sent.

In order to obtain an estimation of the received signal, we probed the knowledge acquired by the learners. Specifically, after the exposure phase of the experiments (Radulescu et al., 2019), in the test phase, participants were asked for grammaticality judgements (yes/no answers) on four types of test strings: Familiar-syllable XXY, New-syllable XXY, Familiar-syllable X1X2Y (i.e. strings of three different but familiar syllables) and New-syllable X1X2Y (i.e. strings of three different and new syllables). These test strings were used as questions to probe the received message, as it was reconstructed by the learner's receiver after decoding the received signal.

Since, as defined above, the message sent was an XXY rule-based language, i.e. strings of 3 syllables with a *same-same-different* pattern, regardless of whether they were familiar or new syllables, correct answers were acceptance of XXY strings, with familiar or new syllables, as possible in the familiarization language, and rejections of X1X2Y strings, with familiar or new syllables. Results showed that the correct acceptance of New-syllable XXY strings increased gradually as the input entropy increased from signal version S1 up to S6, while the correct rejection of Familiar-syllable X1X2Y showed a U-shape pattern. The correct acceptance of Familiar-syllable XXY and correct rejection of New-syllable X1X2Y were consistently high across signal versions. Taken together, these results were interpreted to show that, according to our entropy model, an increase in *input entropy* drives a gradual tendency to move from *item-bound generalization* to *category-based generalization*, since learners were increasingly more likely to accept strings with a *same-same-different* structure, not only with familiar syllables (*item-bound generalization*), but also with new syllables (*category-based generalization*).

From the percentage of correct and incorrect answers to the test items (as a group mean), we can calculate the probability that a signal (y) was received when a signal (x) was sent, i.e. the conditional entropy – Hy(x) – defined by Shannon (1948), as presented above. For example, 73% mean correct acceptance (averaged over the group of participants) of a type of test strings can be interpreted in information-theoretic terms as a $p1 = 0.73$ probability that the intended signal (x) was correctly received, while $p2 = (1 - p1\ ) = 0.27$ is the probability that the incorrect signal (y) was received:

For $p1 \geq p2$:

$Hy(x) = -[p1*logp1 + p2*logp2] = -[0.73 \log(0.73) + 0.27 \log(0.27)] = 0.84$ *b/symbol*

Intuitively, this can be interpreted as the internal entropy of the learner's decision-making system when answering the yes/no question on a type of test strings.

It is important to mention that, since this entropy is based on a behavioral response (i.e. a yes/no answer), it does not constitute a direct measure of the entropy of the received signal. However, it can be envisaged as a *coefficient of equivocation* of the learner when they have to answer the yes/no question based on the signal they received. Recall that rate of equivocation constitutes the bits of information (per second) that are missing in the received signal against the signal sent. Given that an answer to a yes/no question conveys 1 bit of information, this *coefficient of equivocation* gives an estimation of how much information the learners are missing per each bit of information sent by the source every second. Thus, in order to estimate the missing bits of information per second at the receiver's end, i.e. the rate of equivocation, we propose that the entropy per second of the signal sent should be weighted by this *coefficient of equivocation*. For example, at a source rate of information transmission H' = 1.4 bits/s, if the *coefficient of equivocation* is 0.84, the rate of equivocation is estimated at E = 0.84*1.4 = 1.176 missing bits of information per second.

In order to better understand this *coefficient of equivocation,* let us consider the extreme case of nearly total correct acceptance of a type of test strings, that is the nearly perfect case of an ideal rule learner, who correctly accepts a type of test strings in 99% of the cases and only rejects them in 1% of the cases (assuming that the perfect rule learner who accepts these strings 100% of the times can reasonably be considered impossible). In this case, the message is received almost entirely correctly. The *coefficient of equivocation* in this case is $Hy(x) = -[0.99 \log(0.99) + 0.01 \log(0.01)] = 0.07$ *b/symbol*, and thus at the same source rate of transmission of H' = 1.4 b/s, the rate of equivocation E = 0.07*1.4 b/s = 0.1 b/s. Therefore, the actual rate of transmission (R) is the source rate of transmission (1.4 b/s) minus the equivocation rate (0.1 b/s), i.e. the missing bits of information: R = 1.4 – 0.1 = 1.3 b/s, which is a highly efficient rate of transmission, since the missing information from the message sent is nearly zero.

Let us consider now the extreme case of 50% – 50% acceptance vs rejection of a type of test strings (which is considered to be the chance level in behavioral experiments). This means that the message is equally likely to be received as correct or incorrect. The *coefficient of equivocation* in this case is

Hy(x) = – [0.50 log(0.50) + 0.50 log(0.50)] = 1 b/symbol, and thus at the same source rate of transmission of H' = 1.4 b/s, the rate of equivocation E = 1*1.4 b/s = 1.4 b/s. Thus, the actual rate of information transmission is R = H' – E = 1.4 – 1.4 = 0 b/s, so we can say that in this case, in information-theoretic terms, actually no information was transmitted at all: we can obtain the same results by dispensing with the channel completely and just flipping a coin at the receiver's end.

In this way we can calculate the equivocation separately for all test types (Familiar-syllable XXY, New-syllable XXY, Familiar-syllable X1X2Y, and New-syllable X1X2Y), and together they will be a reflection of the total estimated equivocation of the learner when receiving the original signal sent by the source, i.e. for each of the six different versions of the signal (S = {S1, S2, … S6}), to which the learners were exposed in the familiarization.

In the remainder of this section, we will thus show how the total equivocation can be estimated and we will calculate the rate of transmission for all six versions of the signal in the experiments by Radulescu et al. (2019). It must be specified from the beginning that the estimation method used here is not in any case ideal, however it aims at showing in principle how such estimations of rate of equivocation and rate of transmission might be obtained from artificial grammar learning experiments with testing designs based on forced-choice questions. If the testing design includes a production task where participants are asked to produce a list of possible strings in the language (which would be a more natural estimation of the use of a language), the entropy of the produced strings could be directly calculated, and that would be a more straightforward way of estimating the entropy of the received signal, given the sent signal, without having to use a *coefficient of equivocation.*[25]

Since the experiments tested two XXY test types – Familiar-syllable XXY and New-syllable XXY – and two X1X2Y test types – Familiar-syllable X1X2Y and New-syllable X1X2Y – we can calculate a total *coefficient of equivocation* for XXY strings – *H*(XXY) – and a total *coefficient of equivocation* for X1X2Y – *H*(X1X2Y). It makes sense to estimate a total *coefficient of equivocation* for XXY strings and a total *coefficient of equivocation* for X1X2Y strings, since they are closely related in terms of pattern, and thus learners' answers to these types would be highly correlated. In other words, when giving their answers on New-syllable XXY strings and Familiar-syllable XXY strings, the learner would presumably notice the perceptual-identity *same-same-different* pattern, but they would have to decide on the acceptance of familiar and/or new syllables. Similarly, when giving their answers on New-syllable X1X2Y strings and Familiar-syllable X1X2Y

[25] Although in a different type of task, Ferdinand et al. (2018) employ a production task after familiarization, in order to calculate the reduction in entropy of the produced set of items as compared to the familiarization set. We suggest a similar production task should be used in future research, in order to calculate the conditional entropy between the sent signal and the received signal and, thus, obtain a more direct measure of the rate of equivocation.

strings, learners would presumably notice the perceptual *different-different-different* pattern.

It follows that, for *H*(XXY), in the ideal case, if the sent message can be reconstructed with the least errors from the received signal, there will be equal rates of equivocation for the learner when receiving an XXY string, either with familiar or new symbols, and they will both be as close to zero as possible:

H(XXY) = *H*(New-syllable XXY) – *H*(Familiar-syllable XXY) = 0.

For any case which is less than ideal, the received signal will be XXY with more certainty about the Familiar-syllable XXY strings compared to New-syllable XXY strings, since the learner will be able to make a difference between the two types of strings, and this difference will be the total equivocation regarding the XXY strings. More specifically, there will be higher equivocation when receiving New-syllable XXY strings than when receiving Familiar-syllable XXY strings, since the latter strings actually match the strings heard in the familiarization. Following this idea, the equivocation regarding XXY strings would be obtained by subtracting the equivocation for Familiar-syllable XXY from the equivocation for New-syllable XXY. Thus, in these less ideal cases, when the received signal does not match the transmitted signal, the equivocation would be:

H(XXY) = *H*(New-syllable XXY) – *H*(Familiar-syllable XXY) > 0.

In any other case, if there is any other extraneous equivocation, it would be due to external factors, that would impact both *H*(New-syllable XXY) and *H*(Familiar-syllable XXY) equally, such as task challenges, auditory challenges, etc.

Similarly, for *H*(X1X2Y) in the ideal case, if the sent message can be reconstructed with the least errors from the received signal, there will be equal equivocation when receiving an X1X2Y string, regardless of familiar or new symbols, and they will both be as close to zero as possible, with a minimal rate of errors:

H(X1X2Y) = *H*(Familiar-syllable X1X2Y) – *H*(New-syllable X1X2Y) = 0.

In the less ideal cases, there will be less equivocation regarding New-syllable X1X2Y, since this type of strings does not match the familiarized strings in any dimension: unfamiliar syllables and unfamiliar pattern. On the other hand, Familiar-syllable X1X2Y will pose higher uncertainty, since the familiar syllables will match the familiarized strings, but the pattern will not. A less ideal learner, who did not receive the equivocation-free signal, i.e. XXY strings regardless of familiar or new syllables, will make a difference between Familiar-syllable X1X2Y strings and New-syllable X1X2Y strings. Thus, in these less ideal cases, the *coefficient of equivocation* for X1X2Y strings would be:

H(X1X2Y) = *H*(Familiar-syllable X1X2Y) – *H*(New-syllable X1X2Y) > 0.

Since information regarding both XXY and X1X2Y strings is encoded in the same signal (i.e. only XXY strings are present in the familiarization, indirectly implying that X1X2Y are not possible), and there is intercorrelation among the answers for all four test types, because in learner's rationale the answers for XXY would also indirectly be related to the answers for X1X2Y, and vice versa, the average equivocation regarding the received signal can be obtained by averaging over the two *coefficients of equivocation*: *Hy(x)* = avg{ *H*(XXY); *H*(X1X2Y)}.

This average conditional entropy (*coefficient of equivocation*), quantifies the average equivocation per one bit of received signal, and it can be envisaged to reflect the internal entropy of the learner’s decision-making system when answering the yes/no questions based on their received signal.

Using this *coefficient of equivocation*, we can estimate the rate of equivocation. Recall that rate of equivocation constitutes the bits of information that are missing in the received signal (per second) compared to the sent signal. Thus, since the *coefficient of equivocation* quantifies the average equivocation per one bit of received signal, in order to estimate the rate of equivocation associated with the received signal, the source rate of transmission of the signal sent should be weighted by the *coefficient of equivocation*: E = *Hy(x)**H’. Thus, the obtained rate of equivocation constitutes the total number of bits of information (per second) missing from the received signal, namely, the loss of information caused by noise during transmission through the channel.

H	**H'=m*H**	**p1**	**p2**	**p3**	**p4**	**H(XXY new) = – (p1*logp1+ p2*log p2)**	**H(XXY fam) = – (p3*logp3+ p4*logp4)**	**H(XXY)**
		(corr XXY new)	(incorr XXY new)	(corr XXY fam)	(incorr XXY fam)			
2.8	1.40	0.57	0.43	0.95	0.05	0.99	0.29	0.70
3.5	1.75	0.65	0.35	0.98	0.02	0.93	0.14	0.79
4	2.00	0.73	0.27	0.97	0.03	0.84	0.19	0.65
4.25	2.13	0.76	0.24	0.93	0.07	0.80	0.37	0.43
4.58	2.29	0.80	0.20	0.97	0.03	0.72	0.19	0.53
4.8	2.40	0.80	0.20	0.93	0.07	0.72	0.37	0.36

Table 1. Calculations of coefficient of equivocation for XXY strings. Each probability value p = {p1,...p4} was calculated from the percentage of acceptance of the respective test items: e.g. in the experimental condition where participants were exposed to the signal version with H = 4, there was 73% mean correct acceptance (averaged over the group of participants) of New-syllable XXY strings. This percentage can be interpreted in information-theoretic terms as a *p1 = 0.73* probability that the intended signal was correctly received, while *p2 = (1 – p1) = 0.27* is the probability that an incorrect signal was received.

Next, the actual rate of transmission (R = {R1, R2, … R6}) of the message in our experiments from Radulescu et al. (2019) for each signal version (S = {S1, S2, … S6}) would be obtained by subtracting from the source rate of transmission (H’s

= {mH1, mH2, ... mH6}) the total rate of equivocation (E) measured for each group of learners exposed to each signal (E = {S1, S2, ... S6}).

H	**H'=m*H**	**p5**	**p6**	**p7**	**p8**	**H(X1X2Y fam) = –(p5*logp5+ p6*logp6)**	**H(X1X2Y new) = –(p7*logp7+ p8*logp8)**	**H(X1X2Y)**
		(corr X1X2Y fam)	(incorr X1X2Y fam)	(corr X1X2Y new)	(incorr X1X2Y new)			
2.8	1.40	0.83	0.17	0.92	0.08	0.66	0.40	0.26
3.5	1.75	0.91	0.09	0.98	0.02	0.44	0.14	0.30
4	2.00	0.77	0.23	0.97	0.03	0.78	0.19	0.58
4.25	2.13	0.73	0.27	0.82	0.18	0.84	0.68	0.16
4.58	2.29	0.82	0.18	0.93	0.07	0.68	0.37	0.31
4.8	2.40	0.9	0.1	0.83	0.17	0.47	0.66	-0.19

Table 2. Calculations of coefficient of equivocation for X1X2Y strings. Each probability value p = {p5,...p8} was calculated from the percentage of acceptance of the respective test items: e.g. in the experimental condition where participants were exposed to the signal version with H = 2.8, there was 83% mean correct rejection (averaged over the group of participants) of Familiar-syllable X1X2Y strings. This percentage can be interpreted in information-theoretic terms as a *p5 = 0.83* probability that the intended signal was correctly received, while *p6 = (1 – p5) = 0.17* is the probability that an incorrect signal was received.

Tables 1, 2, 3 show the detailed calculations for all the six versions of the signal, and the source rates of transmission.

As can be seen in Fig. 2, the actual rate of transmission (R) increases as a polynomial function (with a polynomial trend analysis showing a nearly-significant quadratic effect (F(2, 3) = 7.881, *p* = .06, R^2 = 0.84) compared to the linearly increasing source rate of transmission (H').

Moreover, the predicted trend in the rate of equivocation (E), i.e. first increasing and then decreasing, is shown in Fig. 3: as the source rate of transmission increases from 1.4 bits/s up to 2 bits/s, the rate of equivocation also increases, but it changes direction into a decreasing trend at a source rate of transmission of 2 bits/s. We deem this change in direction from an increasing to a decreasing trend to indicate that there was indeed an attempt at exceeding the *channel capacity*, which caused an increase in the rate of equivocation. This led to a change in the encoding method, such that a more efficient encoding method was found: the *item-bound generalization* moved *gradually* to the *category-based generalization,* which allows for more and more entropy to be encoded per unit of time with less and less equivocation, that is loss of information, down to an arbitrarily small rate of equivocation – 0.20 bits. With the new encoding method in place (*category-based generalization)*, more data (bits) are being transmitted per unit of time, while the rate of equivocation decreased to a very low rate. This means that the encoding – decoding method is increasingly efficient, such that more bits of information can be transmitted over the channel, while there is less loss of information at receiver's end.

H	**H'=m*H**	**H(XXY)**	**H(X1X2Y)**	**Hy(x)= avg{H(XXY); H(X1X2Y)}**	**E**	**R = H' – E**
2.8	1.40	0.70	0.26	0.48	0.67	0.73
3.5	1.75	0.79	0.30	0.54	0.95	0.80
4	2.00	0.65	0.58	0.62	1.23	0.77
4.25	2.13	0.43	0.16	0.30	0.63	1.50
4.58	2.29	0.53	0.31	0.42	0.96	1.33
4.8	2.40	0.36	-0.19	0.08	0.20	2.20

Table 3. Rate of information transmission in XXY grammar learning

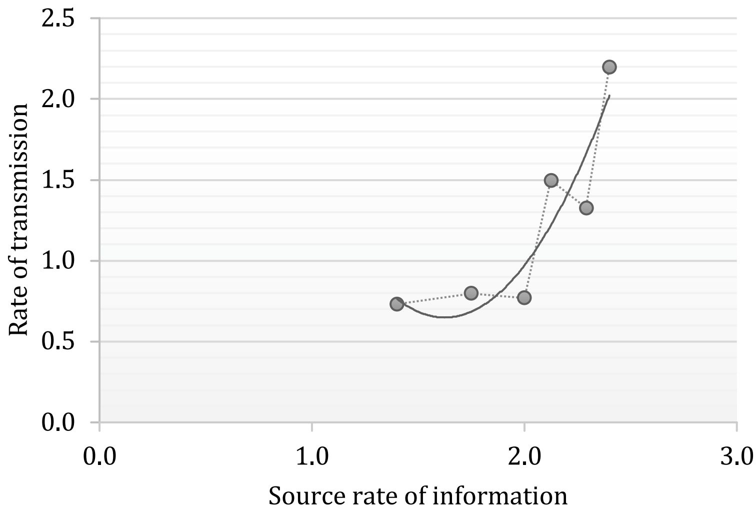


Figure. 2. Rate of information transmission in XXY grammar

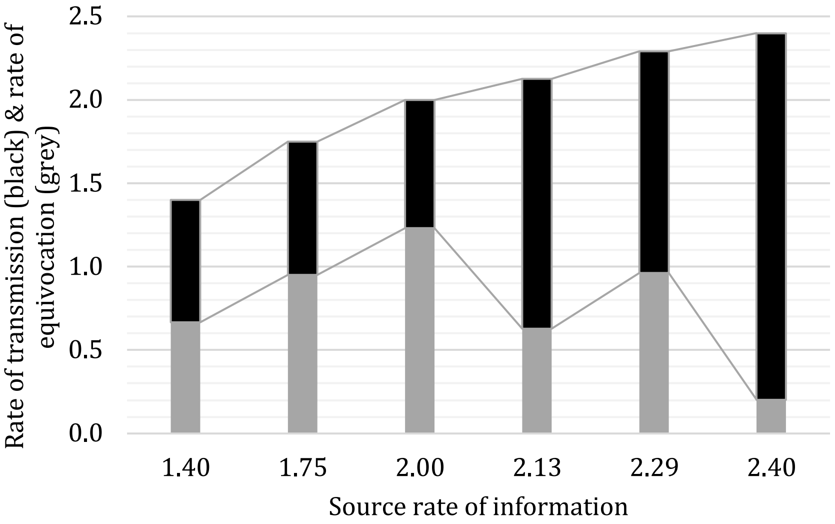


Figure 3. Dynamics of rate of equivocation as a function of increasing source rate of information

Our *channel-capacity*-based prediction that the change in the encoding method is signaled by an increasing trend followed by a decreasing trend of the rate of equivocation, i.e. internal entropy of the learning system, is very much in line with the main tenets of self-organization in the Dynamic Systems Theory: an increase followed by a decrease in system's internal entropy predicts the birth of a new structure (Prigogine & Stengers, 1984; Schneider & Sagan, 2005, Stephen et al., 2009).

To sum up, in this section we gave a brief proof of concept for the *channel capacity* factor in our model, and we proposed an innovative information-theoretical estimation of the rate of transmission and rate of equivocation in an artificial grammar learning task. In addition to the usual data analysis of the group means, the information-theoretical estimations offer a more fine-grained and aggregated insight into the signal received by the learners, i.e. the amount of information received per unit of time and the loss of information per unit of time against the sent signal. We observed an increase followed by a decrease in the rate of equivocation (i.e. missing bits of information per second), caused by an increase in the source rate of information transmission (i.e. an excess of input entropy per second), which is in line with our prediction that *there was **indeed** a change in encoding method which was caused by a **higher source rate of information transmission than the available channel capacity.*** Future research should employ a production task instead of the grammatical judgment test we used in Radulescu et al. (2019), in order to have a more precise and direct quantification of the rate of equivocation, to eliminate the need for a coefficient of equivocation.

The findings of this reinterpretation of the Radulescu et al.'s (2019) results bring strong evidence that indeed the *input entropy per second* was higher than the *channel capacity* in the high entropy versions of the signal (e.g. S6). The next logical step would be to directly speed up the amount of entropy entering the channel per second, i.e. the source rate of transmission (H'), up to the rate from the highest entropy version of the signal (i.e. S6 with H'6 = 2.4 b/s), while keeping the entropy per symbol at the level of the lowest entropy version of the signal (i.e. S1 with H = 2.8 bits). A bit rate which is higher than the *channel capacity* would drive a transition to a more efficient encoding method, i.e. *category-based generalization*.

In the second part of this chapter we present two artificial grammar experiments in which we sped up the source bit rate of information transmission (H'), in order to probe the effect of the time-dependent variable of the *channel capacity* on rule induction.

To the best of our knowledge, these are the first model and the first estimations of rate of information transmission, equivocation rate, and *channel capacity* in an artificial grammar learning task, based on Shannon's noisy-channel theory.

3. Testing the prediction of speeding up the bit rate of information transmission

The second goal of this study is to probe the effect of the time-dependent variable of the second main factor of our entropy model – *channel capacity* – on rule induction, by directly increasing the source rate of transmission (H'), in order to attempt to exceed the *channel capacity*. Theoretically, following the definition of *channel capacity* and Shannon's Theorem 11 (Shannon, 1948), this attempt can be achieved in two ways: either by increasing the amount of entropy (bits) at a constant rate, or by speeding up the rate of feeding information (at constant bit value) into the channel. It follows that, practically, there are two methods to attempt to exceed the *channel capacity*:

1. Add stimulus-unrelated entropy (noise) in the input to render a noisier channel, while keeping the time variable constant. This method aims at exceeding the *channel capacity* by specifically modulating the *noise* variable of the *channel capacity*.

2. Increase the source rate of information production, to directly modulate the time-dependent variable of the *channel capacity*. Specifically, this method aims at reducing the time that the same amount of entropy is sent through the channel, i.e. speeding up the bit rate of information transmission.

We employed the first method in another study (Radulescu, S., Murali, M., Wijnen, F., Avrutin, S., 2021), and we found that added stimulus-irrelevant entropy (background noise), which led to a noisier channel, drove as a consequence a higher tendency towards *category-based generalization*.

Therefore, in this study we employed the second method, i.e. we increased the source rate of information transmission (*input entropy per second*), in order to directly modulate the time-dependent variable of the channel by speeding up its encoding rate. According to the hypothesis of the entropy model, speeding up the source rate of transmission (i.e. to a higher rate than the *channel capacity*) leads to a change in method of encoding, so as to avoid increased rate of equivocation. Why? Because increased rate of equivocation is in fact a loss of information. Thus, the method of encoding transitions to another method of encoding, in order to achieve more efficient transmission of information: that is faster rate of encoding with the least equivocation rate possible. Specifically, we hypothesize that increasing the source rate of information transmission leads to higher tendency to move from *item-bound* to *category-based generalization* for the purpose of achieving a more efficient method of encoding, with the least loss of information (equivocation) possible.

We tested the effect of speeding up the source rate of information transmission on both the repetition-based XXY grammar from Radulescu et al. (2019) study and a more complex grammar – non-adjacent-dependency grammar (aXb). In this type of non-adjacent dependency grammar, specific items *a* always predict specific items *b* over a richer intervening category of *X* items. Learning of such a complex grammar entails both *item-bound generalization* (of the specific items *a* and *b*, and their co-dependency) and also *category-based generalization* of the rich category of intervening *X*s (Gómez, 2002; Grama, Kerkhoff, & Wijnen, 2016; Frost & Monaghan, 2016; Onnis, Monaghan, Christiansen, & Chater, 2004; Wang et al., 2019). This type of artificial grammar learning models the mechanisms needed in language acquisition to acquire rules like *is* go-*ing, is* learn-*ing.* According to our entropy model, a *channel capacity* which is higher than the source rate of information transmission allows for *item-bound generalization*, but, in order to move to a *category-based generalization*, the source rate of information transmission needs to be higher than the channel capacity. So how does the model perform when tested on such a complex type of grammar?

As shown in section 2.4 above, given an entropy (H) of a source and an average number of symbols produced by the source per second (*m*), we can calculate the amount of information produced by the source per second – $H' = mH$ – i.e. the source rate of information transmission. According to Shannon (1948), this amount of entropy determines the *channel capacity* required with the most efficient encoding method, but this entropy cannot exceed the channel capacity. Using this formula, we estimated a source rate of transmission of information in the experiments carried out by Radulescu et al. (2019), as shown in section 2.4 above. Then, we specifically predicted that, if we keep the same information content (input entropy) of the lowest entropy signal version from Radulescu et al. (2019) – where there was no evidence of *category-based generalization*, but we increase the source rate of transmission up to the source rate of transmission of the highest entropy signal version from the same study – where that study found very high tendency towards *category-based generalization*, then we should see a higher tendency to make *category-based* *generalizations*, even though the actual statistical properties (entropy) of the input are the same. Moreover, as we showed in section 2.4, in the highest entropy version of that study we found evidence that indeed the source rate of information transmission was higher than the available *channel capacity,* therefore we wanted to aim for that specific source rate of information.

Specifically, let us denote the source rate of information transmission in the signal version with the highest entropy as $H'_H = m_1H_H$ (1), and the source rate of information transmission in the lowest entropy version as $H'_L = m_1H_L$ (2). Note that the average rate of symbols per second (m_1) was the same in both versions. For the purpose of the manipulation we are aiming for, we would like to obtain $H'_H = H'_L$ but by keeping H_L constant and increasing the average rate of symbols/s to obtain m_2, such that $m_2 > m_1$.

Thus, in the XXY grammar from Radulescu et al. (2019), for a constant *m_1 (symbols/s)*:

H_L= 2.8b/symbol: $H_L' = m_1 H_L$
H_H= 4.8b/symbol: $H_H' = m_1 H_H$

For the purpose of increasing the source rate of transmission up to H_H', while keeping entropy constant (H_L), and by increasing the average rate of symbols/s, we calculated the necessary m_2, as follows:

$m_2 H_L = H_H'$
$m_2 H_L = m_1 H_H$
$m_2/ m_1 = H_H / H_L$
$m_2 = (4.8/2.8)\ m_1$
$m_2 = 1.71\ m_1$

Thus, we obtained $m_2 = 1.71m_1$, and translated it into duration of syllables and within- and between-string pauses, such that we sped up all elements (syllables and pauses) proportionally by a coefficient of 1.71. As a result, we created a faster source rate of information transmission, i.e. entropy per second ($H_L' = H_H'$), but we kept the entropy per symbol constant H_L= 2.8b/symbol.

Next, for the aXb grammar, we created two versions of the signal with different entropy levels (H_L; H_H), but the same average rate of symbols/s (m_3):

H_L= 3.52b/symbol: $H_L' = m_3 H_L$
H_H= 4.71b/symbol: $H_H' = m_3 H_H$

For the purpose of increasing the source rate of information transmission up to H_H', while keeping entropy constant (H_L), and by increasing the average rate of symbols/s, we calculated the necessary m_4, as follows:

$m_4 H_L = H_H'$
$m_4 H_L = m_3 H_H$
$m_4/ m_3 = H_H / H_L$
$m_4 = (4.71/3.52)\ m_3$
$m_4 = 1.34\ m_3$

Thus, we obtained $m_4 = 1.34m_3$, and translated it into duration of syllables and within- and between-string pauses, such that we sped up all elements (syllables and pauses) proportionally by a coefficient of 1.34. As a result, we created a faster source rate of information transmission, i.e. entropy per second ($H_L' = H_H'$), but we kept the entropy per symbol constant H_L= 3.52b/symbol.

In addition to probing the direct effect of the time variable of *channel capacity*, as presented above, this study also looked into the effect of individual differences in cognitive capacities on rule induction, namely the effect of those capacities hypothesized by our entropy model to underlie the *channel capacity*: memory capacity and a domain-general pattern-recognition capacity. To this end, we tested each participant on three independent tests: a Forward Digit Span task, which is a measure of explicit short-term memory (Baddeley et al., 2015), an incidental memorization task, which measures implicit memory capacity (Baddeley et al., 2015), and Raven's Standard Progressive Matrices (RAVENS – Raven et al., 2000), which is a standardized test of fluid intelligence based on visual pattern-recognition (Carpenter et al. 1990, Little et al. 2014). Thus, according to the hypotheses of our entropy model, we predicted a positive effect of RAVENS on the tendency to move from an *item-bound* to a *category-based generalization*, and a negative effect of the explicit/incidental memory tests on the same transition from one type of encoding to the other.

Therefore, we designed and carried out two experiments, one to test the effect of sped up rate of information on rule induction in an XXY grammar, and the other one to test the same effect in an aXb non-adjacent dependency (NAD) grammar. To the best of our knowledge, these are the first language learning experiments that investigate the effect of the time-dependent variable of the *channel capacity* in rule induction, by specifically testing information-theoretic predictions made by an entropy model.

4. Experiment 1

In Experiment 1, participants carried out three tasks. The first task presented the XXY grammar in two different conditions: a slow source rate of information transmission (Slow Rate condition) and a fast source rate of information transmission (Fast Rate condition). In the Slow Rate condition, we used the exact stimuli and source rate of information transmission (H_L') as in the lowest entropy condition from Radulescu et al. (2019) – 2.8 bits. In the Fast Rate condition, the same stimuli were used (H_L = 2.8), but the source rate of information transmission was sped up by a factor of (H_H/H_L=) 1.71 (as per the calculations presented in section 3 above). In the test phases participants were presented with four different types of test strings, just as in the design by Radulescu et al. (2019), which we briefly presented in section 2.4 above, and on which we will elaborate here for further clarification and to formulate our predictions. Participants were presented with a grammaticality judgement task, where they had to answer a yes/no question to indicate whether the test strings could be possible in the familiarization language. The test included four types of test strings, in order to test how the participants encoded the familiarization stimuli, as presented below.

Familiar-syllable XXY (XXY structure with familiar X-syllables and Y-syllables) – correct answer – yes – accept. This type of test strings was used to test learning of the familiar strings. Both groups were expected to accept this type of strings as grammatical, either due to having encoded them as *item-bound* *generalizations* (in the Slow Rate condition), or as *category-based generalizations* (in the Fast Rate condition).

New-syllable XXY (XXY structure with new X-syllables and Y-syllables) – correct answer – yes – accept. This type was used to test whether learners moved from *item-bound generalization* to *category-based generalization* which enables them to accept XXY strings with new syllables. Therefore, we expected that the Fast Rate group was more likely to accept this type of strings as grammatical, as compared to the Slow Rate group. However, absolute mean acceptance rate of this type of strings does not represent direct evidence for category-based generalization. As we argued in Radulescu et al. (2019), this mean should be compared to the mean acceptance rate of Familiar-syllable XXY strings: if the difference of the mean acceptance rate between New-syllable XXY strings and Familiar-syllable XXY strings is significantly smaller in the Fast Rate condition as compared to the Slow Rate condition (i.e. effect size), this would suggest that learners were more likely to have formed *category-based generalization* in the Fast Rate condition than in the Slow Rate condition.

Familiar-syllable X_1X_2Y (X_1X_2Y structure with familiar syllables) – correct answer – no – reject. Participants are expected to confidently reject this type of strings, either by having encoded the input as *item-bound generalizations* (as we expect the Slow Rate group) or *category-based generalizations* (the Fast Rate group). Specifically, participants in the Slow Rate condition are expected to confidently reject this type of strings, as their memory trace of the Familiar-syllable XXY strings is expected to be strong enough to highlight a mismatch between these strings and the Familiar-syllable X_1X_2Y strings. Participants in the Fast Rate condition are expected to form strong *category-based generalizations*, thus they should confidently reject the Familiar-syllable X_1X_2Y strings as deviant from the *same-same-different* rule.

New-syllable X_1X_2Y (X_1X_2Y structure with new syllables) – correct answer – no – reject. Participants are expected to confidently reject this type of strings, either by having encoded the input as *item-bound generalizations* (as we expect the Slow Rate group) or *category-based generalizations* (the Fast Rate group).

The second task was a Forward Digit Span, which is a standard measure of short-term memory capacity (Baddeley et al., 2015). The third task was an incidental memorization task, which measures the ability to memorize information without being explicitly instructed to do so (Baddeley et al., 2015). According to the hypotheses of our entropy model, we predicted a negative effect of the explicit/incidental memory capacities on learners' tendency to move from an item-bound encoding to a category-based encoding.

Importantly, we tested the same participants in both experiments, which were conducted in two separate sessions, on two different days (at least three days passed between sessions). For practical reasons, all participants took part first in the aXb grammar experiment (Experiment 2) and then in the XXY grammar experiment (Experiment 1). For theoretical presentation reasons, that have to do with the logic and theoretical development of the entropy model and its hypotheses, here we present the XXY experiment first, followed by the aXb experiment.

4.1 **Participants**

Fifty-six adults, Dutch native-speakers (10 male, age range 18-72, M_{age} = 26.39, SD_{age} = 11.06) participated. All participants were naïve to the aim of the experiment and had no known language, reading or hearing impairment or attention deficit. Participants received 5 euros for their participation in Experiment 1. One additional participant was tested, but excluded after having reported to suffer from Attention Deficit Disorder.

4.2 Materials

Task 1: XXY grammar

Familiarization stimuli. Participants listened in both the Slow Rate and the Fast Rate conditions, to the same XXY artificial grammar used in the low entropy condition of Experiment 2 in the study by Radulescu et al. (2019). The three-syllable strings of the language display an XXY structure (each letter stands for a set of syllables). Each string consists of two identical syllables (XX) followed by another different syllable (Y): e.g. *ke:ke:my, da:da:li*. All syllables consist of a consonant followed by a long vowel, to resemble common Dutch syllable structure. The subset of X-syllables does not overlap with the subset of Y-syllables. Overall, seven X-syllables and seven Y-syllables were used to generate seven strings (see Appendix A for complete stimulus set). Each string was repeated four times in each familiarization phase (7 strings x 4 repetitions = 28 strings in each familiarization phase).

The same 28 strings were used for all three familiarization phases, such that the entropy was the same in all familiarization phases – 2.8 bits. For the entropy calculations, we employed the same method as in Radulescu et al. (2019), which is a fine-tuned extension of a related entropy calculation method proposed by Pothos (2010) for finite state grammars (see Table 4 below for complete entropy calculations).

The order of presentation of the strings was randomized for every participant, and each participant was randomly assigned to either the Slow Rate or the Fast Rate condition, in order to obtain a between-subjects experimental design. In the Slow Rate condition there was a pause of 50 ms between the syllables within strings, and a pause of 750 ms between the strings. In the Fast Rate condition all X and Y syllables, as well as the within-and between-string pauses were sped up separately by a factor of 1.71, using Praat 6.0.49 (64-bit Edition for Windows; Boersma & Weenick, 2005).

Low Entropy

H[bX]=H[7] = -Σ[0.143*log0.143] = 2.8
H[XX] = H[7]= 2.8
H[XY] = H[7] = 2.8
H[Ye] = H[7] = 2.8
H[bXX] = H[7] = 2.8
H[XXY] = H[XYe]= H[7] = 2.8
H[bigram] = 2.8
H[trigram] = 2.8
H[total] = $\frac{\mathbf{H[bigram]+H[trigram]}}{\mathbf{2}}$ **=** 2.8

Table 4. Entropy value for Experiment 1. Taken from Radulescu et al. (2019)

Test stimuli. In total there were three familiarization phases, interleaved with three (quick) intermediate test phases and a final (longer) test phase. Each intermediate test phase included four test strings, one of each type. The final test had eight test strings (two of each type). Thus, in total, there were (4+4+4+8=) 20 test strings (see Appendix A for the complete set of stimuli). Accuracy score for the learning of the XXY grammar was measured as correct acceptance of Familiar-syllable XXY and New-syllable XXY strings and correct rejection of Familiar-syllable X_1X_2Y and New-syllable X_1X_2Y strings.

We recorded all the yes/no answers and coded them as correct acceptance of Familiar-syllable XXY and New-syllable XXY strings and correct rejection of Familiar-syllable X_1X_2Y and New-syllable X_1X_2Y strings. From all the 20 correct/incorrect answers for each participant we calculated a proportion of correct answers per each type of test item. Next, instead of directly analyzing proportions, we performed an empirical logarithmic transformation, in order to analyze the data using a linear model.

Task 2: Forward Digit Span

Participants were explicitly told that this was a memory test, during which a series of digits would be presented aurally, and they would have to recall them in the same order. To prevent participants from creating a visual pattern on the keypad while listening to the digits, we modified the standard Forward Digit Span task such that no physical keyboard was made available to the participants, rather a row with buttons for each digit was displayed in a line on the screen only in the moment when they were asked to enter the digits by clicking the buttons, and disappeared during the listening phases. We used the standard scoring method: we measured the highest span of each participant and recorded it as one data point per participant.

Task 3: Incidental Memorization Test

Participants listened to 30 bisyllabic nonsense words resembling Dutch phonology. Crucially, participants were not told in advance that a memory test would be administered. They were only told that they were about to listen to words from another forgotten language. They were instructed to imagine what the word might have meant in the forgotten language and to pick a category (flower, animal, or tool), based on what the word sounded like to them. They had 3 seconds to choose a category for each word, by pressing the button for flowers, animals, or tools.

After this phase, a message informed the participants that they would be given a memory test, which would check whether they remembered the words they categorized during the previous phase. They were instructed to press a yes/no button on the screen, depending on whether they heard the word previously or not. In the memorization test participants gave answers on 13 targets and 13 foils. We recoded all correct/incorrect answers into a *d'* value for each participant.

4.4 Procedure

Participants completed the tasks in the order presented above. For Task 1, they were told that they would listen to a "forgotten language" that would not resemble any language they might know, which had its own rules and grammar. Participants were informed that the language had more words than what they heard in the familiarization phases. They were told that each intermediate test would be different from the other tests, and the tests were meant to check what they had noticed about the language. They had to decide, by pressing a Yes or a No button, if the words they heard in the tests could be possible in the language. This task lasted around 5 minutes. For Task 2, they were explicitly instructed that it was a memory test. For Task 3, they were not told in advance about the memory test. The entire experiment lasted about 20 minutes.

4.5 Experiment1: Results

Figure 4 presents the mean correct acceptance rate (proportion of correct acceptances per group) for Familiar-syllable XXY strings and New-syllable XXY strings, across the two conditions (Slow Rate, Fast Rate). The mean correct acceptance rate in the Slow Rate condition for Familiar-syllable XXY strings was *M* = .96 (*SD* = .1), and for New-syllable XXY strings it was *M* = .75 (*SD* = .27). The mean rate of correct acceptance in Fast Rate condition for Familiar-syllable XXY strings was *M* = .99 (*SD* = .04), and for New-syllable XXY strings it was *M* = .9 (*SD* = .18).

Similarly, Figure 5 shows the mean correct rejection rate (proportion of correct rejections per group) for Familiar-syllable X_1X_2Y strings and New-syllable X_1X_2Y strings, across the Slow Rate and Fast Rate conditions. In the Slow Rate condition, the mean correct rejection rate for Familiar-syllable X_1X_2Y

strings was *M* = .93 (*SD* = .24) and for New-syllable X_1X_2Y strings was *M* = .99 (*SD* = .04). In the Fast Rate condition, the mean correct rejection rate for Familiar-syllable X_1X_2Y strings was *M* = .99 (*SD* = .05), and for New-syllable X_1X_2Y strings was *M* = .99 (*SD* = .08).

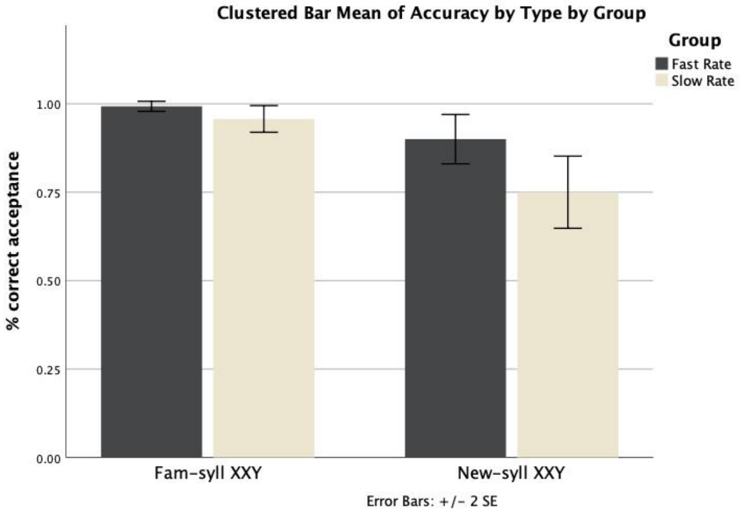


Figure 4. Mean rate of correct acceptance for Familiar-syllable XXY and New-syllable XXY strings in both conditions: Fast Rate and Slow Rate. Error bars show standard error of the mean.

In order to probe the effect of *channel capacity* on rule induction, we compared the performance in the two conditions (Slow Rate and Fast Rate groups) in a general linear mixed effects analysis of the relationship between Accuracy (correct acceptance of the grammatical test items and correct rejection of the ungrammatical ones) and the Rate of Transmission (Slow Rate, Fast Rate) as well as the Type of Test Strings (Familiar-syllable XXY, New-Syllable XXY, Familiar-syllable X_1X_2Y, New-Syllable X_1X_2Y). Therefore, as dependent variable we entered Accuracy score into the model. As fixed effects we entered Rate of Transmission, Type of Test Strings and Rate of Transmission x Type of Test Strings interaction. As random effect we had intercepts for subjects. The scores for Forward Digit Span, Incidental Memorization Task and RAVENS tests[26] were entered one by one as covariates in the model. 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. The model reported here is the best fitting model, both in terms of the model's accuracy in predicting the observed data, and in terms of AIC (Akaike Information Criterion).

[26] RAVENS scores were obtained for the participants during the second experiment presented in this paper, since the same participants participated in both experiments (see section 5 below).

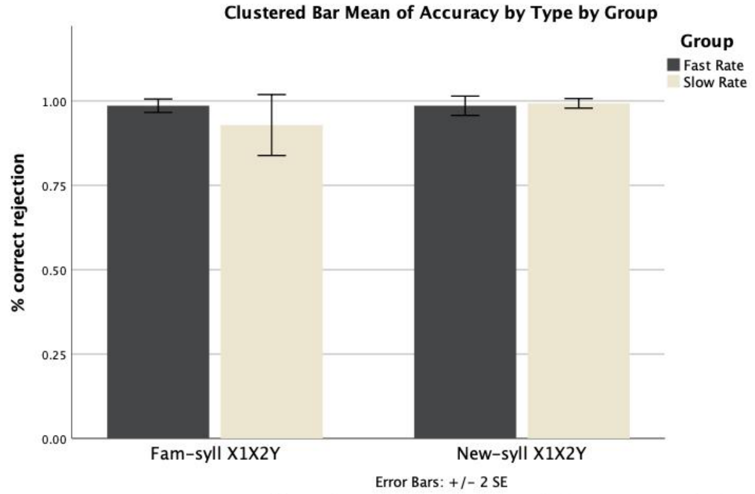


Figure 5. Mean rate of correct rejection for Familiar-syllable X1X2Y and New-syllable X1X2Y strings in both conditions: Fast Rate and Slow Rate. Error bars show standard error of the mean.

We found a significant main effect of Type of test strings (F(3, 213) = 5.742, *p* = .001), a marginally significant Rate of Transmission x Type interaction (F(4, 213) = 2.039, *p* = .090), a non-significant Forward Digit Span effect (F(1, 213) = .069, *p* = .793), a non-significant Incidental Memorization Task effect (F(1, 213) =.880, *p* = .349) and a non-significant RAVENS effect (F(1, 213) = 2.326, *p* = .129).[27]

Pairwise comparisons of the Estimated Marginal Means (adjusted to the mean values of the covariates in the model, i.e. Forward Digit Span = 6.68, Incidental Memorization Task = 1.968, RAVENS = 71.54) revealed a significant difference between the Rate of Transmission conditions (Fast Rate and Slow Rate groups) for the New-syllable XXY (M = .101, SE = .045, F(1, 213) = 4.936, *p* = .027), and a nearly-significant difference for the Familiar-syllable X1X2Y (M = .085, SE = .045, F(1, 213) = 3.522, *p* = .062). For the other two Types of test, pairwise comparisons of the Estimated Marginal Means adjusted for the same level of the covariates revealed a non-significant difference between the Rate of Transmission conditions (Fast Rate and Slow Rate groups): Familiar-syllable XXY (M = .010, SE = .045, F(1, 213) = .051, *p* = .822) and New-syllable X1X2Y (M = .012, SE = .045, F(1, 213) =.069, *p* = .793).

Furthermore, Cohen's effect size value (d) and the effect-size correlation (r) for the difference in acceptance between Familiar-syllable XXY and New-syllable XXY was higher in the Slow Rate condition (d = 1.03, r = 0.45; large effect size), than in the Fast Rate condition (d= .69, r = .32; medium effect size).

[27] We also checked the main effect of Rate of Transmission, and since it was non-significant (F(1, 213) = 2.558, *p* = .111), it did not improve the model, and it created effects of an overfitted model, we excluded it from the final model presented here.

In information-theoretic terms, the rate of equivocation (E) dropped from 0.60 bits/s down to 0.46 bits/s in the Fast Rate group, i.e. when the source rate of transmission (H') was sped up from 1.40 bits/s up to 2.39 bits/s, while the input entropy was kept constant at 2.8 bits/symbol. As a consequence, the actual rate of transmission (R) increased significantly from 0.8 bits/s up to 1.93 bits/s (Table 5 shows the relevant calculations based on the formulas from section 2.4).

H	**Source rate (H')**	**H(XXY)**	**H(X1X2Y)**	**Hy(x)= avg{H(XXY); H(X1X2Y)}**	**E**	**R = H' – E**
	(m*H)					
2.8	1.40	0.57	0.29	0.43	0.60	0.80
2.8	2.39	0.39	0.00	0.19	0.46	1.93

Table 5. Rate of information transmission and rate of equivocation in the Slow Rate condition (H' = 1.40) vs. Fast Rate condition (H' = 2.39)

4.6 Discussion

The results of Experiment 1 show that the mean acceptance of new XXY strings as grammatical in the familiarization language was higher in the Fast Rate condition than in the Slow Rate condition, as predicted by our model. Moreover, there was a difference between the rates of acceptance of new XXY strings vs. familiar XXY strings depending on the rate of transmission group: there was a smaller difference between the mean acceptance of the new XXY strings vs. familiar XXY strings in the Fast Rate condition compared to the Slow Rate condition. This shows differences between groups in terms of how they encoded the input: if learners do not make a clear distinction between a new XXY string and a familiar XXY string, we conclude that they encoded the input as *category-based generalization,* which allows them to accept both a new and a familiar XXY string based on a *same-same-different* rule regardless of new or familiar syllables. Hence, a smaller difference between the means of acceptance of these test types in the Fast Rate condition shows a higher tendency towards *category-based generalization* than in the Slow Rate condition. Thus, these results together show that there was a higher tendency towards *category-based generalization* when the source rate of transmission was sped up to a higher rate than the *channel capacity*, even though the input entropy was the same in both conditions, which supports the predictions of our entropy model regarding the effect of the time-dependent variable of the *channel capacity* on rule induction.

The rate of correct rejection of X1X2Y strings with familiar syllables was also higher in the Fast Rate condition than in the Slow Rate condition, which supports the same hypothesis of our model: when the source rate of transmission was sped up, learners formed *category-based generalizations* which helped them reject strings that were deviant from the *same-same-different* rule, although they had familiar syllables.

In information-theoretic terms, the results of this experiment show that speeding up the source rate of information transmission caused the transition to a more efficient encoding method, which is signaled by the significant drop in the rate of equivocation: while the source rate of transmission increases from 1.4 bits/s up to 2.39 bits/s, the rate of equivocation drops from 0.60 bits/s down to 0.46 bits/s. This shows that there was indeed an attempt at exceeding the *channel capacity*, which caused a change in the encoding method, such that a more efficient encoding method was found: the *item-bound generalization* transitioned to the *category-based generalization,* which allows for more entropy to be encoded per unit of time with less equivocation, that is less loss of information at receiver's end. With the new encoding method in place, more data (bits) are being transmitted per unit of time, while the rate of equivocation (i.e. loss of information) decreases to a very low rate. This indicates that the encoding – decoding method is more efficient, such that more bits of information can be reliably transmitted over the channel, that is a higher actual rate of information transmission is attained, while there is less loss of information when receiving the message (i.e. the XXY language).

When comparing the results of this experiment with the results of the experiments in Radulescu et al. (2019), the actual rate of transmission from the Fast Rate group of this experiment (R = 1.93 bits/s) is almost as high as the actual rate of transmission from the highest entropy condition (H = 4.8 bits/symbol) from Radulescu et al. (2019): R = 2.20 bits/s), although in the experiment of the present study the input entropy was just as low as the lowest entropy condition from Radulescu et al. (2019), i.e. H = 2.8 bits/symbol. This shows that, even at a low input entropy, speeding up the source rate of information transmission drives a change of the encoding method towards a more efficient encoding, which allows for higher rate of information transmission with less equivocation rate. In other words, the same transition to a more efficient encoding method – *category-based generalization* – was obtained either by increasing the input entropy (H) in Radulescu et al. (2019) or by reducing the time that the same input entropy is fed into the channel, i.e. by speeding up the source bit rate of information transmission.

5. Experiment 2

In Experiment 2, participants carried out three tasks. In Task 1, adults were exposed to an *aXb* language where they had to learn item-bound dependencies between *a* and *b* (*item-bound generalization*), while also generalizing *a_b* dependencies over a category of *X* words (*category-based generalization*). For example, they had to learn the item-bound dependency *tɛp_jɪk*, and generalize it over new *X* elements (like *nilbo, perxɔn*): *tɛp_nilbo_jɪk, tɛp_perxɔn_jɪk,* etc.

We designed two experimental conditions: a slow source rate of information transmission (Slow Rate condition) and a fast source rate of information transmission (Fast Rate condition). As presented in section 3, we first created two entropy versions of the grammar, but the same average rate of symbols/s (m_3), then we increased the average rate of symbols/s (m_4), in order to reach the same source rate of information transmission of the high entropy version, but, crucially, keeping the input entropy low.

Unlike Gómez (2002), we kept *X* set size constant (18 *Xs*) and manipulated entropy by combining each of the three *a_b* frames with different subsets of 6 *Xs* (3 *a_b* * 6 *Xs*) which generated a rather low entropy signal (H_L = 3.52 bits/symbol). For the high entropy signal, the *aXb* grammar combined exhaustively each of the three *a_b* frames with all intervening *Xs* (3 *a_b* * 18 *Xs*), which resulted in a rather high entropy (H_H = 4.7 bits/symbol). Since such evaluations of low/high entropy could be seen as relative, depending on the grammar/language, we took into account previous studies on non-adjacent dependency learning (Gómez, 2002; Grama, Kerkhoff, & Wijnen, 2016; Radulescu & Grama, 2020), in order to estimate the set size and variability of such an *aXb* language that would be necessary to achieve a low entropy version and a high entropy version. For the entropy calculations, we employed the same implementation model as in Radulescu et al. (2019) – see Table 6 below for complete entropy calculations.

Thus, we obtained the following:

H_L= 3.52 b/symbol: $H_L' = m_3 H_L$

H_H= 4.71 b/symbol: $H_H' = m_3 H_H$

In the Slow Rate condition, we used the low entropy version as presented above H_L= 3.52b/symbol: $H_L' = m_3 H_L$. In the Fast Rate condition, the same stimuli were used (H_L= 3.52b/symbol), but the source rate of information was sped up by a factor of (H_H/H_L = 4.71/3.52 =) 1.34 (as per the calculations presented in section 3 above).

Low Entropy	**High Entropy**

H[begin-*a*]=H[3] =
-Σ[0.333*log0.333] = 1.58
H[aX] = H[18] = 4.17
H[Xb] = H[18] = 4.17
H[*b*-end] = H[3] = 1.58
H[begin-aX] = H[18] = 4.17
H[aXb] = H[Xb-end] = H[18] = 4.17
H[bigram] = 2.86
H[trigram] = 4.17
H[total] = $\frac{\mathbf{H[bigram]+H[trigram]}}{\mathbf{2}}$ **=** 3.52
	H[begin-*a*]=H[3] =
	-Σ[0.333*log0.333] = 1.58
	H[aX] = H[54] = 5.75
	H[Xb] = H[54] = 5.75
	H[*b*-end] = H[3] = 1.58
	H[begin-aX] = H[54] = 5.75
	H[aXb] = H[Xb-end] = H[54] = 5.75
	H[bigram] = 3.67
	H[trigram] = 5.75
	H[total] = $\frac{\mathbf{H[bigram]+H[trigram]}}{\mathbf{2}}$ **=** 4.71

Table 6. Entropy values for the two entropy versions of the *aXb* grammar

In the test phase, participants were asked to give grammaticality judgements on aXb strings with either correct (familiar) or incorrect (unfamiliar) *a_b* frames. Importantly, all (correct and incorrect) test strings included new X elements which were not present in the familiarization. If learners correctly accept *aXb* strings with the correct *a_b* frames and new *X* elements, it shows they were both able to encode *item-bound generalizations* (i.e. the *a_b* frames) and to generalize them over a category of *X* elements, i.e. *category-based generalization*.

Recall that, according to our entropy model, rule induction is an encoding mechanism that *gradually* goes from *item-bound generalization* to *category-based generalization* as a function of the interaction between the input entropy (and more specifically, the *source rate of information transmission*) and the *channel capacity*. In other words, this is a phased mechanism that goes gradually from the first phase of *item-bound generalization* to the next-level phase of *category-based generalization*.

Learning of *aXb* strings requires both *item-bound generalization* of the *a_b* frames simultaneously with *category-based generalization* of these frames over a category of *X* elements. In such a case a sped up source rate of information transmission attempts to exceed the channel capacity and drives the transition to *category-based generalization* faster, such that the item-bound encoding mechanism for the *a_b* frames might be phased out, and the encoding method might move to *category-based generalization* for the *a_b* frames as well, not only for the *X* category of intervening elements. Specifically, this means learners might encode the *a* and *b* elements, as categories, which does not restrict the dependencies to only between a specific *a* element and a specific *b* element. That is learners might not encode an $a_i\_b_i$ relationship, but a relationship between a category of *a* elements and a category of *b* elements, which allows also for an $a_i\_b_j$ dependency to be legit ("class-words" – Endress & Bonatti, 2007). To sum up, the predictions for the outcome of this task could be actually opposites for the two types of relationships encoded in such an *aXb* grammar: speeding up the source rate of information transmission attempting to exceed channel capacity impedes *item-bound generalization* (of the specific $a_i\_b_i$ relationship), but it facilitates *category-based generalization* (i.e. generalizing a relationship between *a* and *b* categories over a category of *Xs*).

The second task participants had to complete was RAVENS Standard Progressive Matrices, (Raven et al., 2000). According to the hypotheses of our entropy model, we predicted a positive effect of RAVENS on the tendency to move from *item-bound* to *category-based generalization*.

In the third task, participants completed a word-recall task, designed to test item memorization, i.e. detailed phonological representations of the *a*, *b* and *X* elements, in order to test for a correlation between learners' representations of specific items and their accuracy scores. We expected accurate memorization of the *a*/*b* elements to support better learning of the *a_b* dependencies, and thus better accuracy scores. Conversely, failing to recall *X*s would indicate better generalization of the *X* category, hence better scores.

5.1 Participants

The same 56 participants from Experiment 1 participated in Experiment 2. We tested one more participant in Experiment 2 (as Experiment 2 was conducted prior to Experiment 1, one participant did not return to participate in Experiment 1), so this participant was excluded from the analysis. Therefore, in total 57 participants took part in Experiment 2 (10 male, age range 18-72, M_{age} = 26.28, SD_{age} = 11), and received 10 euros for their participation.

5.2 Materials

Task 1: aXb grammar learning

Familiarization stimuli. All *a* and *b* elements were monosyllabic nonsense words (e.g., *tɛp, jɪk*), while all *X* elements were bisyllabic nonsense words (e.g., *naspu, dyfo:*). Each *a_b* pair was combined with a different, non-overlapping set of 6 *X* elements (see Appendix B for the complete stimuli set). In both Slow Rate and Fast Rate conditions, two versions of the *aXb* language were used: Language 1 (L1) and Language 2 (L2). The only difference between L1 and L2 was the specific legit combination of the three *a* and *b* elements into pairs: *tɛp _lyt, sɔt_ jɪk* and *rak_tuf* (L1), and *tɛp _ jɪk, sɔt_tuf* and *rak_lyt* (L2). Therefore, every a_i _b_i pair in L1 was ungrammatical ($a_i\_b_j$) in L2, and vice versa. The reason for two different versions was to prevent an effect of idiosyncrasies of particular *a_b* combinations (L1 or L2). Therefore, each version of the *aXb* grammar (L1, L2) consisted of (3 $a_i\_b_i$ * 6 X_i =) 18 different $a_iX_ib_i$ strings. Each participant was exposed to only one version of the *aXb* grammar (either L1 or L2), and to only one source rate of transmission condition (either Slow Rate or Fast Rate).

The 18 different $a_iX_ib_i$ strings were presented 12 times, resulting in a total of 216 strings, in a randomized order for each participant. In the Slow Rate condition there was a 100ms within-string pause, and a 750ms between-string pause. In the Fast Rate condition all the *a, b* and *X* elements, as well as the within-string and between-string pauses for each *aXb* string were sped up by a factor of 1.34 (as per calculations in section 3). To this goal, we used Praat 6.0.49 (64-bit Edition for Windows, Boersma and Weenick, 2005). As in Experiment 1, the duration of each *a, b* and *X* word was shortened separately by the 1.34 factor, using the "Duration Factor" argument of the "Change Gender" command, and then the elements were spliced into the specific *aXb* strings.

Test stimuli. Each *a_b* frame of each language (L1, L2) was combined with two novel *X* elements to yield (6 *a_b* * 2 *X* =) 12 new test items (see Appendix B for the test *X* elements). Hence, each participant was exposed to 12 new *aXb* strings, six of which were grammatical and six ungrammatical. The six new *aXb* strings which contained the L1 *a_b* pairs were counted as ungrammatical for the L2 learners, while the six new *aXb* strings with the L2 *a_b* pairs were ungrammatical for the L1 learners. Accuracy scores for learning the *aXb* grammar were calculated as correct acceptances of the grammatical test strings and correct rejections of the ungrammatical test strings.

Task 2: RAVENS

The second task was Raven's Standard Progressive Matrices (Raven et al., 2000), for which participants had to solve 60 matrices, by identifying which pattern is missing in a multiple-choice task. Each matrix consists of a set of nine patterns arranged in a particular order according to some underlying rules, of which one pattern is missing. The standard RAVENS allows 50 minutes for completion, but, after a pilot, we allowed participants only 35 minutes, to avoid a time-consuming and exhausting experiment session. We used the standard scoring method: we counted all correct answers, and then we used the standard tables to transform them into age-corrected percentiles.

Task: 3: Word-recall task

The Word Recall task had two tests. In the first test, participants were presented visually with 12 familiar 2-syllable *X* words from the *aXb* language, and 12 new bisyllabic foils, similar to the familiar ones, which overlapped in one syllable with the target words. The second test presented participants visually with 6 monosyllabic familiar *a* or *b* elements of the *aXb* language, and 6 new nonsense word foils, which differed from the target words only by one letter (see Appendix C for stimulus set). Participants had to indicate for each word, whether they had heard it during the first task. Accuracy scores were measured as correct acceptances of the familiar items and correct rejections of the foils.

5.3 Procedure

Before the familiarization phase of Task 1 participants were instructed that they would listen to an "alien language" that does not resemble any language that they might be familiar with, and which has its own rules and grammar. To avoid any motivation to explicitly look for patterns in the stimuli, participants were not informed of the subsequent test phase until after the end of the familiarization phase. Before the test phase, participants were instructed that they would listen to new sentences in the same "alien language", none of which would be identical to the sentences they had heard before. They were then asked to decide for each sentence whether it was correct or not, according to the grammar of the language they had just heard, by clicking on "Yes" or "No". They were instructed to answer quickly and intuitively. Afterwards, the other tasks were administered in the order from above. Experiment 2 lasted approximately one hour.

5.4 Results

Table 7 shows the means and standard deviations of accuracy scores (proportion correct responses) for both conditions (Slow Rate vs Fast Rate).

Condition	*M*	*SD*	*n*	*SE*	*95% CI for Mean Difference*
Slow Rate	0.69	0.46	29	0.09	0.51, 0.87
Fast Rate	0.55	0.50	28	0.09	0.37, 0.74

Table 7. Descriptive statistics of mean correct score in two conditions of exposure. Experiment 2

Figure 6 shows boxplots of the distribution of individual mean accuracy rates (correct acceptances/rejections) in each condition, i.e. Slow Rate and Fast Rate.

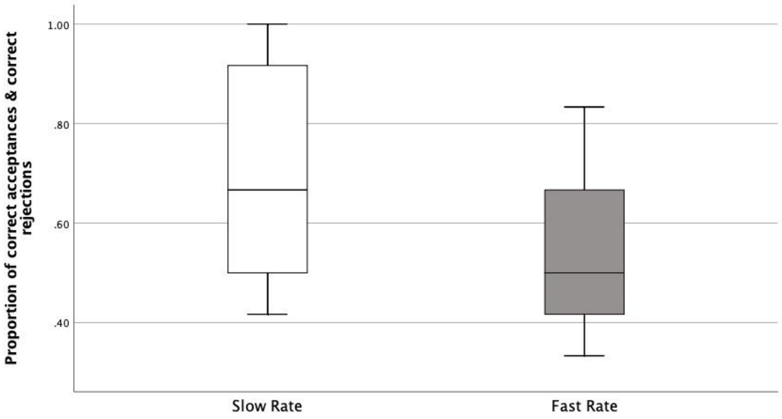


Figure 6. Boxplots of proportions of correct acceptances and correct rejections (accuracy rate) in Slow Rate condition as compared to Fast Rate condition

Figures 7 and 8 show the histograms of individual mean accuracy scores (correct acceptance/rejection) in Slow Rate condition and in Fast Rate condition, respectively. Specifically, Figure 7 shows a bimodal distribution of individual accuracy scores in the Slow Rate condition: this shows that most participants either performed around chance level or achieved a very high accuracy score. Figure 8 shows most participants in the Fast Rate condition performed between 40% and 60%.

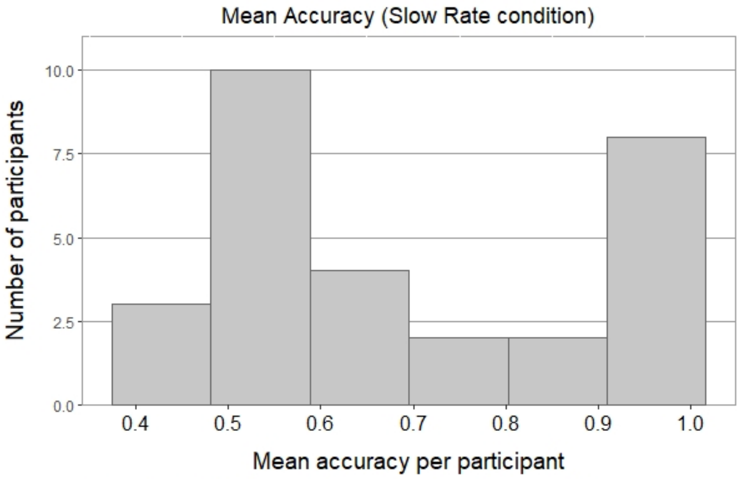


Figure 7. Histogram of mean accuracy per participant in Slow Rate condition.

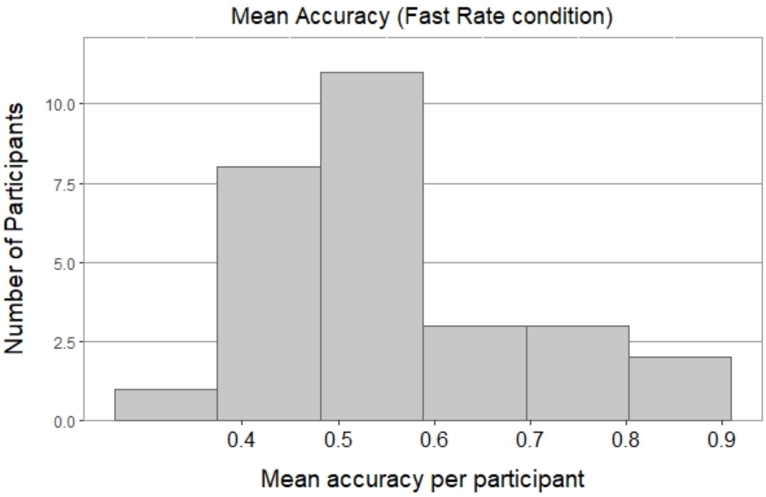


Figure 8. Histogram of mean accuracy per participant in Fast Rate condition.

Because the data was not normally distributed, a non-parametric statistical test was used to assess whether learning performance in Fast Rate condition was significantly different from chance level. To this end, a two-tailed one-sample Wilcoxon signed-rank test was performed. Accuracy score of participants in the Fast Rate condition (M = .55, SD = .50) was found to be significantly different from chance level at the .05 level of significance, with a moderate effect size (p = .017, 95% CI for mean difference .5 to .63, r = .45). Accuracy score of participants in the Slow Rate condition (M = .69, SD = .46) was found to be significantly different from chance level at the .05 level of significance, with a large effect size (p < .001, 95% CI for mean difference .67 to .83, r = .73).

To compare performance across the two conditions we used R (R Core Team, 2017) and the *glmer* function of the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017) to perform a general linear mixed effects analysis of the relationship between Accuracy (correct acceptance of grammatical test strings and correct rejection of ungrammatical test strings) and the Rate of Transmission (Slow Rate, Fast Rate). As dependent variable we entered Accuracy in the model, and as fixed effects we entered Rate of Transmission (Slow Rate, Fast Rate) and Language (L1, L2), without interaction term. As random effects we had intercepts for subjects and items. 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. The model reported here is the best fitting model, both in terms of model's accuracy in predicting the observed data, and in terms of AIC (Akaike Information Criterion). Likelihood Ratio Tests were performed separately as a means to attain p-value for the effect of each predictor (Rate of Transmission, Language).

A significant main effect of Rate of Transmission ($\chi^2(1)$ = 8.64, p = 0.003, conditional R^2 = .13) on Accuracy was found. These results show that participants who were exposed to the Fast Rate of transmission of the aXb grammar had significantly lower Accuracy scores as compared to participants who were exposed to the Slow Rate of transmission of the same aXb grammar. Language was not a significant predictor ($\chi^2(1)$ = 3.2, p = 0.07, conditional R^2 = .13. Finally, no significant interaction effect was found between Rate of Transmission and Language ($\chi^2(1)$ = .29, p = 0.6, conditional R^2 = .13). The scores of individual differences tests (Forward Digit Span, Incidental Memorization Test, Raven's Progressive Matrices, Word Recall Test) were added to this model as fixed factors, one by one. None of the individual differences tests significantly improved the model when added as fixed factors one by one in the model, however only the accuracy score in the Word Recall Test for a / b (but not X) elements of the *aXb* grammar had a significant positive effect on the Accuracy scores ($\chi^2(1)$ = 3.8, p = .05, conditional R^2 = .1).

5.5 Discussion

In Experiment 2 we tested the effect of speeding up the source rate of transmission on learning a complex *aXb* grammar, which requires both *item-bound generalization* of the specific *a_b* dependencies, and *category-based generalization* in order to generalize those dependencies over a category of intervening *X* elements. According to our entropy model, our predictions for this experiment were opposite for the two types of relationships encoded in an a_iXb_i grammar: increasing the source rate of information transmission impedes *item-bound generalization* (of the specific $a_i\_b_i$ relationship), but it facilitates *category-based generalization* (i.e. generalizing a relationship between *a* and *b* categories over a category of *Xs*).

The results showed that there was indeed a significant effect of increasing the source rate of transmission on learning the *aXb* grammar, such that the Fast Rate group scored lower than the Slow Rate group. This shows that increasing the source rate of transmission by a factor of 1.34 in this particular a_iXb_i grammar with an entropy of 3.52 bits/symbol makes learning of the specific $a_i\_b_i$ frames and generalizing them over novel intervening *X* elements more difficult than a slower rate of transmission. Moreover, participants who recalled the *a/b* elements better across conditions learned the specific $a_i\_b_i$ frames better. Thus, learning of a_iXb_i grammar is correlated with item-specific encoding of *a/b* elements. These results support the predictions of our entropy model, namely that an increased source rate of information transmission impedes *item-bound generalization* (of the specific $a_i\_b_i$ relationship).

As we argued above, if learners correctly accept new *aXb* strings with the specific familiar $a_i\_b_i$ dependencies and new *X* elements, it shows they were both able to encode *item-bound generalizations* ($a_i\_b_i$ frames), and to generalize them over a category of *X* elements, i.e. *category-based generalization*. This is what happened both in the Slow Rate and the Fast Rate condition. However, the Fast Rate group had a lower tendency to do so compared to the Slow Rate group. There could be several logical interpretations: either Fast-Rate learners failed at *category-based generalization* of the *Xs,* or they failed at *item-bound generalization* of the $a_i\_b_i$ frames, or they were simply confused. Therefore, we looked into the acceptance/rejection ratios. If the first case was true, rejection rates should be higher than acceptance rates, since all test items had new *Xs*. This was not the case. Actually, Fast-Rate learners show similarly high acceptance rates for both language-specific a_iXb_i strings (specific to the exposure language, e.g. L1) and language-deviant a_iXb_j strings (specific to the other language, e.g. L2), with a rather high acceptance rate for the language-deviant a_iXb_j strings (Median=.58) compared to the Slow-Rate learners (Median=.33) (Figures 9 and 10).

This points to the fact that the Fast-Rate learners failed to learn the specific $a_i\_b_i$ dependencies, that is *item-bound generalization* was impaired in the Fast Rate group. If this was the case, this result can be accounted by our entropy model, as we argued in section 5, a sped up source rate of information transmission precipitates the transition to *category-based generalization* faster, such that the item-bound encoding mechanism for the specific $a_i\_b_i$ frames might be phased out, and the encoding method moves to *category-based generalization* for the $a_i\_b_i$ frames as well. This would be a case of overgeneralization: categories of *a/b* elements would be inferred (i.e. *category-based generalization*), not just the item-bound specific $a_i\_b_i$ frames, so any *a* could freely combine with any *b*, such that $a_i\_b_j$ frames would also be accepted ("class-words"). Since all test items show new combinations with *X* elements, the learner might find it highly probable that the *a/b* elements could yield new combinations, as long as they preserve the main *aXb* order and word characteristics (i.e. monosyllabic *a* followed by a bisyllabic *X* and then a monosyllabic *b*).

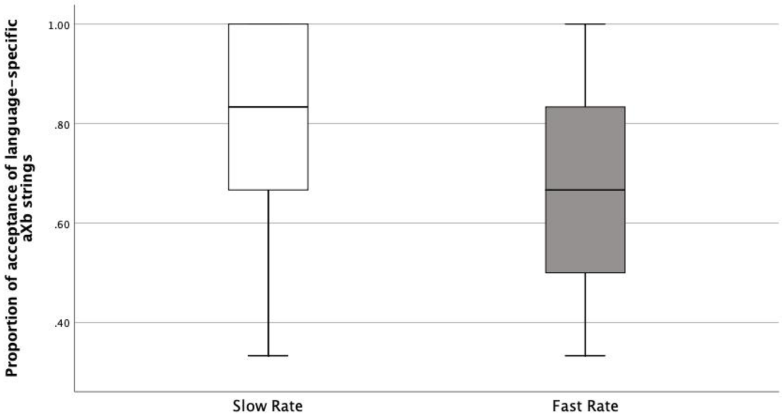


Figure 9. Boxplots of proportions of acceptance of language–specific aXb strings in Slow Rate condition as compared to Fast Rate condition

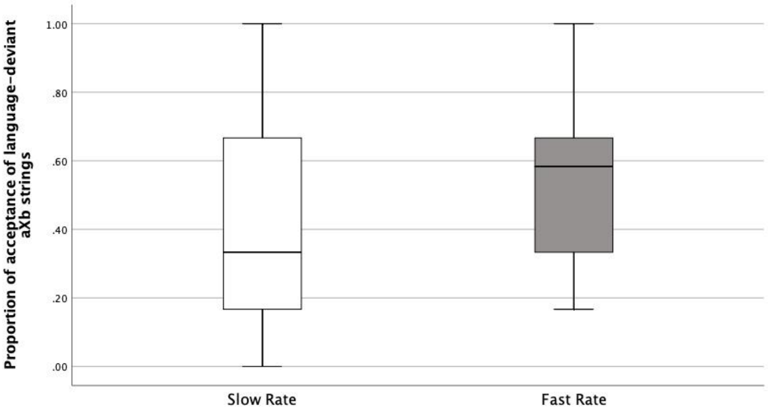


Figure 10. Boxplots of proportions of acceptance of language–deviant aXb strings in Slow Rate condition as compared to Fast Rate condition

Following this logic, if Fast-Rate learners actually overgeneralized, they must have started the test by accepting both language-specific and language-deviant *aXb* strings, and after the first acceptances they would question why all the test items seem to be acceptable, which might have led to an increased rate of rejections in the last part of the test. Alternatively, if Fast-Rate learners were just confused, the acceptances should be randomly scattered over test trials.

An inspection of the acceptance rate of both language-specific and language-deviant *aXb* strings, in the Fast Rate condition, showed a higher tendency to accept all the test strings in the first three trials of the test (*t(11)* =

-1.951, *p* = .05), regardless of exposure language, than in the last trials. These results might point to a case of overgeneralization in the Fast Rate condition.

Thus, it is possible that the source rate of information transmission was increased to a higher extent than required to actually learn the a_iXb_i grammar and it led to overgeneralization. Further research should specifically test the overgeneralization hypothesis, and further look into the effect of sped up source rate of information transmission at a lower rate, i.e. a speeding up factor $m < 1.34$, to find the adequate source rate of transmission for learning this complex grammar.

6. General Discussion and Conclusions

Understanding the cognitive mechanisms involved in learning not only the item-bound regularities in the input, but also in the emergence of new categories and structures has been a major topic of research in cognitive science. This paper contributes to the ongoing research and debate on the underlying mechanisms and the factors that drive both *item-bound* and *category-based generalization*, by further extending the entropy model for rule induction that we proposed in Radulescu et al. (2019), which offers a more refined formal approach to the classical *Less-is-More hypothesis* (Newport, 1990; 2016). According to this hypothesis and other related studies (Gerken, 2006; Gómez, 2002; Hudson Kam and Newport, 2005, 2009; Hudson Kam and Chang, 2009; Reeder et al., 2013), rule induction was either deemed to be driven by statistical properties of the exposure language, i.e. input variability, or by limitations of the cognitive capacities involved in this process, i.e. memory capacities. However, *how* and *why* these two factors should play a role in rule induction, and the exact cognitive capacities and mechanisms that lead to the emergence of new structures remain largely unexplained.

Our entropy model (Radulescu et al., 2019) took a step further by bringing together these two factors in one information-theoretic account based on Shannon's noisy-channel coding theory (Shannon, 1948), and proposed that rule induction is an encoding mechanism resulting from the interaction between the variability of the input, i.e. *input entropy,* and the finite time-sensitive encoding capacity of our brain, which we envisage as the *channel capacity,* in information-theoretic terms. Specifically, our model hypothesizes that an increase in the *input entropy per second* which attempts to exceed the finite *channel capacity* drives the transition from *item-bound generalization* to *category-based generalization*. In Radulescu et al. (2019), in two artificial grammar experiments that tested the effect of the first factor, *input entropy*, we found evidence that supports our model, namely that when input entropy increases, the tendency to move from *item-bound generalization* to *category-based generalization* increases gradually.

However, our model specifically predicts that it is not high entropy in absolute terms which is the factor at stake in this mechanism, but it is our time-sensitive entropy-processor – *channel capacity* – which places an upper bound on the amount of entropy that is needed per unit of time in order for the encoding mechanism to move from a method of encoding to another. Therefore, this study had two goals: theoretical, to further specify and refine our entropy model for rule induction that we proposed in Radulescu et al. (2019), by showing a concrete example of how we can apply Shannon's *channel capacity* and his noisy-channel coding theory (Shannon, 1948) to rule induction in an artificial grammar learning; and experimental, namely we directly manipulated the time-dependent variable of the *channel capacity* in two other artificial grammar experiments, in order to specifically probe the essential effect of *channel capacity* on rule induction as hypothesized by our entropy model.

Firstly, we showed how the encoding mechanism of rule induction in an artificial grammar learning environment can be modeled in terms of Shannon's communication system theory, and we offered a brief proof of concept regarding the effect of *channel capacity* on rule induction based on the results of our previous experiments reported in Radulescu et al. (2019). Specifically, based on the experimental data we obtained in that study, in this study we showed how we can calculate and estimate the following information-theoretic measures, which are key to rule induction, according to our entropy model: the source rate of information transmission, the rate of equivocation and the maximum rate of information transmission, i.e. *channel capacity*.

More precisely, we showed how we can probe experimentally the specific prediction we derived from Shannon's theory: if indeed the source rate of information transmission (*input entropy per unit of* time) is higher than learners' available *channel capacity,* then the transition from one encoding method to a more efficient encoding method as hypothesized by our model, i.e. from *item-bound* to *category-based generalization,* should be signaled by an initial increase followed by a decrease of the rate of equivocation (i.e. missing bits of information). And indeed results showed that an increase in the source rate of information transmission caused an initial increase followed by a decrease of the rate of equivocation. This shows that in order to cope with a higher inflow of entropy per unit of time, the encoding system found a new encoding method which allows for its maximum rate of information transmission to be reached: it moved from an inefficient encoding method (i.e. with high loss of information at receiver's end), to a more efficient encoding method, which allows for higher input entropy to be encoded reliably (with the least loss of information possible) per unit of time. This finding is in accord with the main tenets of Dynamic Systems Theory, according to which an increase followed by a decrease in system's internal entropy predicts the birth of a new structure (Prigogine & Stengers, 1984; Schneider & Sagan, 2005, Stephen et al., 2009).

In conclusion and to the best of our knowledge, this is the first study that shows an innovative way to calculate and measure experimentally the increase and decrease of the equivocation rate (i.e. loss of information at receiver's end) in order to estimate the *channel capacity*, and to show (in information-theoretic terms) the transition from *item-bound generalization* to *category-based generalization* in artificial grammar learning.

As to the second goal of our study, in two artificial grammar experiments, we probed the effect of *channel capacity* on rule induction by directly manipulating the time-dependent variable of the *channel capacity*, i.e. we sped up the source rate of information transmission. According to the entropy model, an increase in the source rate of information transmission which is higher than the *channel* capacity, drives a higher tendency to move from an *item-bound generalization* to a *category-based generalization*. We probed this hypothesized effect on rule induction in two types of artificial grammars: an XXY grammar, and a more complex aXb grammar.

Learning of the XXY grammar requires learners to abstract away from specific items of the X and Y categories, and to move from an *item-bound generalization* to a *category-based generalization,* that is to learn a *same-same-different* rule between categories, regardless of specific items included in these categories. Results showed that this transition from one encoding method to the other was driven by an increase of the source rate of information transmission, i.e. *the input entropy per unit of time*, while the statistical properties of the input, i.e. *input entropy per symbol,* remained constant at a very low level. This result showed that indeed, as hypothesized by our entropy model, rule induction is an encoding mechanism that moves from *item-bound generalization* to *category-based generalization* as a result of the interaction between the *input entropy* and our time-sensitive entropy-processor *channel capacity.*

Next, we employed a more complex *aXb* grammar in order to pose a challenge to the model: learning of this type of *aXb* grammar requires learners to move from an *item-bound* to a *category-based generalization* for the *X* category of middle elements, while, crucially, to stick to an *item-bound generalization* for the specific *a_b* dependencies. If increased source rate of information transmission drives *category-based generalization* for the *X* category, it follows that it should impede *item-bound generalization* for the specific *a_b* dependencies of such an *aXb* grammar. Indeed, the results showed that faster source rate of information caused a lower accuracy than slower source rate of information on this type of grammar, that is when exposed to a faster rate of the *aXb* grammar, learners failed to generalize the specific *a_b* dependencies over new intervening *X* elements. In accord with our model, one possible interpretation of these results would be that the source rate of transmission was too high for this type of grammar with the specific input entropy that we tested (3.52 bits), and thus it precipitated the transition to *category-based generalization* for the specific *a_b* dependencies as well, not only for the middle *X* elements. This points to a case of possible overgeneralization, where learners might have learned an *AXB* grammar, where A and B also stand for categories, instead of *item-bound* relations between specific *a* and *b* elements over a category of *X* elements. Indeed, it is possible that for this type of grammar fast, but not furious, might yield better learning. Future research should look into a slower rate of information transmission for an *aXb* grammar with this specific entropy (3.52 bits).

The effect of time on generalization behavior has only been marginally investigated in cognitive sciences: a few studies used other ways than information-theoretic approaches in order to investigate the effect of a time-dependent variable on category learning (Reeder, et al., 2009; 2013), in non-adjacent dependency learning (Endress & Bonatti, 2007; Wang et al., 2016; 2019) and in auditory statistical learning (Emberson et al., 2011). Converging evidence from all these studies, despite using different approaches to the temporal variable (exposure time, speech rate, temporal distance between stimuli), highlights a clear pattern: generally a shorter time is beneficial to auditory rule (category) learning. There is also some converging evidence from neural networks research that reduced training time results in model's better generalization, or lower generalization error (Hardt, Recht, & Singer, 2016). Our study contributes to this research topic by taking a step further and applies a purely information-theoretic measure directly derived from Shannon's noisy-channel coding theory and based on the quantified amount of input entropy per second (bits/s) of the signal. In this sense, our model and findings offer a more principled and fine-tuned approach to our time-sensitive entropy processor involved in rule induction.

Although in a different domain of application, our model is also compatible with another information-theoretic hypothesis derived from Shannon's noisy-channel coding theory – namely, the hypothesis of Uniform Information Density (Jaeger, 2006; Levy & Jaeger, 2007; Jaeger, 2010) – which proposes that in language production speakers prefer (intuitively) to encode their message by a uniform distribution of information across the signal, with a rate of information transfer close to the channel capacity, but without exceeding it. In other words, language production is inherently a mechanism designed for efficient communication, in that it balances the amount of information per time or per signal (dubbed "information density"), in such a way that the channel is never under- or overutilized (Jaeger, 2010). According to this hypothesis, underutilization means a waste of channel, while overutilization brings the risk of information loss, as per Shannon's noisy-channel coding theory, and therefore according to the Uniform Information Density. By posing the noisy-channel capacity as an upper bound of the rate of information transmission for the purpose of efficient transmission without loss of information, our model accounts for the Uniform Information Density hypothesis, and takes a step further by offering a more general domain of application (i.e. learning and generalization), and a more refined way to quantify the rate of information transmission and estimate *channel capacity*.

At the algorithmic level (in the sense of Marr, 1982), our entropy and channel capacity model for rule induction in artificial grammar is compatible with recent models of recognition memory (Cox & Shiffrin, 2017) and exemplar models applied to artificial grammar learning (Jamieson & Mewhort, 2010). Future research should look into the link between our entropy model and these formal approaches based on encoding instances as vectors of features, with generalization being triggered by vector similarity (Chubala & Jamieson, 2013). Indeed, as we argued in Radulescu et al. (2019), by refining the feature similarity approach to category formation proposed by Aslin & Newport (2012; 2014), our entropy model suggests that information is re-structured from *item-bound* to *category-based generalization* by (unconsciously) re-observing the structural properties of the input and identifying similarities (shared features) and specific differences (unshared features) between items. Crucially, our model proposes the *channel capacity* as the upper bound on the amount of similarities/differences encoded. The degree of specificity of the encoding (i.e. *item-bound* specificity) is given by the amount of differences encoded with specific items, which results from a lower or higher *input entropy* (measured in bits of information): the more differences are encoded (higher *input entropy*), the higher the degree of specificity of the encoding (i.e. *item-bound generalization*). Conversely, when the degree of specificity of the encoding reaches the upper bound placed by the *channel capacity* on the number of bits encoded per second, a reduction or "gradual forgetting" of the encoded differences is triggered, in order to avoid an inefficient, i.e. noisy, encoding (Radulescu et al., 2019). Hence, more and more similarities between items are highlighted, which drives an automatic gradual grouping of items under the same "bucket". Hence, the degree of specificity decreases and the degree of generality increases *gradually* with each bit of information. Thus, a gradient of specificity/generality on a continuum from *item-bound* to *category-based generalizations* can be envisaged in terms of number of bits of information encoded in the representation (analogous to the degree of stability/plasticity in terms of strength of memory pathways in neural networks – Abraham & Robins, 2005).

A follow-up research question would be to better define and specify the *channel,* whether it is a communication channel between speakers, or an abstract channel as we mostly hinted in this study, i.e. an abstract channel between an abstract source – a grammar – and a learner. However, in this study we also briefly suggested a more in-depth and granular understanding of the abstract concept of *channel* as a system of channels: intuitively, and oversimplifying here, the signal from the environment enters learner's acoustic channel, which has a specific rate of information transmission, then the output of this channel becomes the input to the perception channel, whose output becomes the input to the cognitive channel. Estimates of the bit rate of information processing by applying information theory were proposed in some perception and cognitive domains, e.g. in visual attention (Vergese & Pelli, 1992), in visual processing (Koch et al., 2006), unconscious vs conscious processing (Dijksterhuis & Nordgren; 2006), cognitive control (Wu et al., 2016). However, we suggest that the concept of *channel* should be first and foremost defined and specified in physical and biological terms (i.e. at the level of brain structure and neural networks), and further investigated in terms of its link to the cognitive capacities (at the algorithmic level). That would mean further investigating and applying Shannon's *channel* and *noisy-channel coding theory* to recent developments in neurobiology, where it was shown that artificially-induced forgetting at the cellular level drives generalization (Migues et. al, 2016). Moreover, since information is physical (Karnani, Pääkkönen, & Annila, 2009; Laughlin, de Ruyter van Steveninck, & Anderson, 1998, Machta, 1999), further research should look into the information-theoretic concept of *channel* and *rate* *of information transmission* at the level of neural pathways. The neural pathways are the physical/biological medium (i.e. channel) transmitting one form of information (acoustic energy) to the brain, i.e. encoded and decoded, into another form of information (i.e. neuronal energy – patterns of electric activity at the neuronal level). The physical bioprocesses of energy transformation from acoustic information into electric signal and transmission through neural networks have been proposed to underlie abstract memory representations (Collell & Fauquet, 2015; La Cerra, 2003; Varpula, Annila, & Beck, 2013).

Before concluding, it is imperative to clarify one aspect. A model of *finite* and *noisy*-channel capacity might lead the reader to assume a kind of a cognitive limitation as in a flaw of the cognitive system, which is definitely not the case. We do not propose a model in which the emergence of rules and categories, i.e. structure, is merely the side effect of some constraints of a limited biological system. In accord with innovative theories and findings in neurobiology (Frankland, Köhler, & Josselyn, 2013; Hardt, Nader, & Wang, 2013; Migues et. al, 2016; Richards & Frankland, 2017), we deem our *finite* and *noisy*-channel capacity to be a design feature of our biological system for adaptive purposes. More precisely, neurobiological evidence shows that our memory system is designed to encode memories not as in-detail representations of the past, but as simplified models better suited for future generalization in noisy environments (Richards & Frankland, 2017). The brain employs several strategies to undermine faithful in-detail representations to prevent overfitting to past events (in accord with neural networks research – Hawkins, 2004; MacKay, 2003), which promotes better generalization (among which, *noise* injection – a neurobiological mechanism that increases random variability in the synaptic connections – Villarreal et al., 2002).

Fast but not furious, reads the title of this article. Speed up, but not wildly and in an unrestrained fashion. The channel capacity acts as the speedometer, and determines the maximum rate of information transmission with the adequate encoding. In this study, we proposed an innovative method to increase the rate of information to tax *channel capacity*. We found that increasing the rate of transmission by a specific factor calculated by applying Shannon's formula to experimentally obtained data indeed has the hypothesized effect on rule learning: it drives *category-based generalization,* and it interferes with *item-bound generalization*. Thus, we deem necessary to specify that by sped up bit rate we do not mean that an unrestrained increased bit rate, in absolute terms, up to very high bit rates drives rule induction in any context, or grammar. In other words, the very specific dynamics between the *input entropy* and the maximum *rate of information transmission* drive rule induction. Further research should investigate this sweet-spot and find the mathematical relation between these two factors.

Appendix A

Familiarization strings
keːkeːmy
jujuɣo
daːdaːli
pypyveː
tøːtøːrøː
hihisaː
fofoʃu
keːkeːmy
jujuɣo
daːdaːli
pypyveː
tøːtøːrøː
hihisaː
fofoʃu
keːkeːmy
jujuɣo
daːdaːli
pypyveː
tøːtøːrøː
hihisaː
fofoʃu
keːkeːmy
jujuɣo
daːdaːli
pypyveː
tøːtøːrøː
hihisaː
fofoʃu

Test strings

Test 1		**Test 2**	**Test 3**	**Final Test**	
Familiar-syllable XXY	daːdaːli	hihisaː	keːkeːmy	tøːtøːrøː	jujuɣo
New-syllable X1X2Y	poxaːru	runyni	xaːmisy	syniny	mininy
New-syllable XXY	dydytaː	zuzuvo	sosory	jijiføː	ʋuʋuseː
Familiar-syllable X1X2Y	judaːsaː	pytøːmy	keːfoveː	hidaːrøː	tøːpyɣo

Appendix B

	a/b	***IPA***
a1	tep	[tɛp]
a2	sot	[sɔt]
a3	rak	[rɑk]
b1	lut	[lyt]
b2	jik	[jik]
b3	toef	[tuf]
	X	
No.	***Familiarization***	***IPA***
1	blieker	[blikər]
2	dufo	[dyfo]
3	fidang	[fidɑŋ]
4	gopem	[xopəm]
5	kengel	[kɛŋəl]
6	kijbog	[kɛibɔx]
7	loga	[loxa]

8	malon	[malɔn]
9	movig	[movix]
10	naspu	[nɑspu]
11	nijfoe	[nɛifu]
12	noeba	[nuba]
13	plizet	[plizɛt]
14	rajee	[raje]
15	rogges	[rɔxəs]
16	seeta	[seta]
17	snigger	[snixər]
18	wabo	[vɑbo]
	Test (novel Xs)	
19	nilbo	[nilbo]
20	pergon	[perxɔn]

Appendix C – *Word Recall X*

Target X word used	*X word not used*	*Foil*	*Word Recall a/b*	*Target*	*Foil*
[movix]	[nɑspu]	[nɑsfu]		[tɛp]	[fɛp]
[vɑbo]		[lɛipu]		[sɔt]	[sɔs]
[nɛifu]	[xopəm]	[xobər]		[rɑk]	[rɑuk]
[seta]		[vapəm]		[lyt]	[lym]
[kɛŋəl]	[raje]	[rafo]		[jik]	[juk]
[rɔxəs]		[poje]		[tuf]	[xuf]

[dyfo]	[blikər]	[blifot]			
[nuba]		[prukər]			
[kɛibɔx]	[loxa]	[lopɛi]			
[snixər]		[pixa]			
[fidɑŋ]	[malɔn]	[mazət]			
[plizɛt]		[silɔn]			

Chapter 6

Turn That Noise On. Noisy Backgrounds Drive Rule Induction

Radulescu, S., Murali, M., Wijnen, F., and Avrutin, S.[28]

Abstract

Forming general representations from exposure to a limited set of specific examples, in other words, the emergence of structure, has been a long-standing hot topic for research in cognitive sciences. However, the underlying mechanism and the necessary environment for these abilities to take shape remain largely underspecified. Here we further extend and test our information-theoretic model (Radulescu et al., 2019; 2021) for rule induction, according to which a higher *input entropy* than the available encoding capacity (*channel capacity*) drives the tendency to move from a high-specificity *item-bound generalization* to another more general form of encoding, *category-based generalization*. In this study, we further tested the model by looking into the effect of a *noisy-channel capacity* (as defined by Shannon, 1948), to attempt to exceed the encoding capacity (*channel capacity*) by adding *noise* (i.e. random stimulus-irrelevant material) in the background of an artificial language learning task. Specifically, while exposing adults to an XXY artificial grammar, we played random digits and beeps in the background, in order to create a *noisy environment.* In one condition learners had to pay attention and remember specific digits from the noise material, while participants in another condition were not given any additional task on the background noise material. We found that added signal-irrelevant entropy (noise) drives the tendency towards *category-based encoding*, regardless of the low target-intrinsic entropy in the input, but crucially only when no additional task was required on the noise material. The *noisy-channel capacity* at the computational level maps onto what can be envisaged as an attentionally-taxed and error-prone encoding system with time-dependent limitations at the algorithmic level, and not to an overloaded task-handler. To the best of our knowledge this is the first artificial grammar experiment that investigates the effect of *noisy-channel capacity* on rule induction, by specifically testing information-theoretic predictions made by an entropy model in order to disentangle the effect of *noisy-channel capacity* from the effect of overloading the

[28] This chapter is a modified version of a manuscript in preparation:
Radulescu, S., Murali, M., Wijnen, F., & Avrutin, S. (2021) Turn That Noise On. Noisy Backgrounds Drive Rule Induction

underlying cognitive capacities with additional tasks. Our findings show that specific noisy environments drive rule induction, in accordance with the information-theoretic approach on *noisy-channel capacity,* and with dynamic systems theory, where noise is a catalyst for self-organizing into new structures (Stephen et al., 2009).

1. Introduction

Rule induction (generalization or regularization) was hypothesized and shown to result either from processing and encoding the variability in the input (Gerken, 2006; Reeder, Newport, & Aslin, 2013), or from overloading certain limited cognitive capacities, e.g. memory capacity (Hudson Kam & Chang, 2009; Hudson Kam & Newport, 2009; Newport, 1990). This ability ranges from finding statistical patterns between specific items in the input (Saffran, Aslin, & Newport, 1996; Thiessen & Saffran, 2007) to a more abstract learning that allows for category/rule induction (Marcus et al, 1999; Smith & Wonnacott, 2010; Wonnacott, 2011; Wonnacott & Newport, 2005). However, the underlying mechanism and the exact dynamics between these triggering factors for the inductive step from memorizing specific items and statistical regularities to inferring abstract categories/rules remain largely underspecified.

While supporting a *single-mechanism hypothesis*, which was previously proposed to underlie both item-specific and abstract learning (Aslin & Newport, 2012; 2014; Frost & Monaghan, 2016), in a couple of previous studies – Radulescu, Wijnen, & Avrutin (2019) and Radulescu, Kotsolakou, Wijnen, Avrutin, and Grama (2021) – we took a step further in answering the remaining questions. Unlike earlier studies, Radulescu et al. (2019) suggest that, the underlying processes (i.e. the learning mechanisms – *statistical learning* and *abstract rule learning)* should be conceptualized separately from their outcomes, that is from the resulting forms of encoding (*item-bound generalizations* and *category-based generalizations*). While *item-bound generalizations* describe relations between specific physical items, e.g. a relation based on physical identity, like “*ba* follows *ba*”, *category-based generalizations* are operations beyond specific items that describe relations between categories (variables), e.g. “Y follows X”, where Y and X are variables taking different values (for example, *“ba”, “mi”,* etc.).

In order to explain *how* and *why* a single mechanism outputs these two qualitatively different forms of generalization, Radulescu et al. (2019) proposed an entropy model that specifies the dynamics between the two main factors hypothesized by the authors to drive both *item-bound generalizations* and *category-based generalizations:* the statistical properties of the input, i.e. *input entropy*, and the brain’s ability to encode the input under conditions of finite encoding capacity (i.e. *channel capacity*). More specifically, we proposed an entropy model with the main hypothesis that rule induction is an encoding mechanism driven as a natural automatic reaction of the brain due to its sensitivity to the *input entropy* and its finite encoding capacity (*channel capacity*). We define our encoding capacity as *channel capacity,* in information-theoretic terms, which means the finite rate of information encoding (entropy per unit of time), which might be supported by certain cognitive capacities, e.g. memory capacity, in psychological terms.

In this model, external factors (properties of the input) and internal factors (information encoding capacity) interact to drive rule induction: little input entropy facilitates finding rules between specific items, i.e. *item-bound generalization*, while an *input entropy* which is higher than the *channel capacity* drives *category-based generalization*. In two artificial grammar experiments designed to test the first factor, i.e. input entropy – quantified using the formula proposed by Shannon (1948), Radulescu et al. (2019) found evidence that supports the entropy model: when input entropy increases, the tendency to move from *item-bound generalization* to *category-based generalization* increases gradually. These findings bring more granular evidence for the *gradient of generalization* proposed before by Aslin & Newport (2014).

In Radulescu et al. (2021), we probed the effect of the other factor of the model, i.e. the finite *channel capacity*, which is the maximum rate of information encoding (bits/second). Specifically, we manipulated the *time variable* of the *channel* capacity by increasing the speed of information transmission (bits/s). This increase in speed was shown to be higher than the learners' capacity to encode information (*channel capacity*), while keeping *input entropy* at a low level, i.e. the lowest entropy level from Radulescu et al. (2019), where we found no evidence for *category-based generalization*. Results showed an increased tendency towards *category-based generalization* under conditions of sped up inflow of information, even though the *input entropy* was very low.

Taken together, the findings of these two studies (Radulescu et al., 2019; 2021) bring strong evidence to our entropy model, which specifies the quantifiable dynamics between statistical properties of the input (*input entropy*) and properties of the brain's encoding capacity (*channel capacity*): rule induction is driven either by an increase in *input entropy* which is higher than *channel capacity,* or by increasing the speed of the inflow of information up to a higher rate than the *channel capacity.*

In information-theoretic terms, the *channel capacity* was defined in Shannon's noisy-channel coding theory (Shannon, 1948), for communication systems. In simple words, this coding theory says that a message (i.e. information) can only be transmitted reliably (i.e. with the least uncertainty when receiving the message), if encoded by using an efficient encoding method such that the rate of information transmission (bits/second), plus the *noise*, is below the channel's capacity. If the rate of information transmission is higher than the *channel capacity*, then another more efficient encoding method can be found, but the *channel capacity* cannot be exceeded. Since the *channel capacity* is conceptualized as an upper bound on the rate of information transmission, plus the *noise*, it follows that an increase in noise should boost the transmission up to the maximum rate, i.e. *channel capacity*, which in turn precipitates a change in encoding method. Based on these concepts, our model adds into the rule induction "formula" the crucial dimension of *noise*, i.e. random perturbations that interfere with the signal, thus rendering a *noisy channel*. Hence, our model hypothesizes that a noisy environment drives rule induction, that is a change in encoding method, from *item-bound* to a more efficient encoding method, *category-based generalization*. In this study, we further investigate the *channel capacity*, by looking into its *noise* variable and its link to the underlying cognitive capacities involved in rule induction.

Our proposal that rule induction is driven not only by external factors, like input variability, but also by internal factors, like the relevant cognitive capacities involved in processing and encoding information is closely related to another line of research – the classical *Less-is-More* hypothesis (Newport, 1990), which looks into rule induction in terms of cognitive constraints on learning. According to this hypothesis, overloading our limited memory capacity leads to difficulties in storing and retrieving low-frequency items, which prompts overuse of more frequent forms leading to overgeneralization. These maturational limitations on cognitive capacities were proposed to explain the differences in tendency to generalize between young and adult learners (Hudson Kam & Newport, 2005; Hudson Kam & Newport, 2009; Newport, 1990). The findings of this line of research are relevant from a developmental point of view, i.e. the developmental differences between young and adult learners, but they are also relevant to the research topic of this paper, in that they point to the mechanism and the cognitive capacities underlying the *channel capacity* involved in rule induction. Therefore, in the remaining part of this introduction we briefly review previous studies from this line of research with the purpose of identifying possible cognitive capacities underlying *channel capacity*, and in the next section we present in detail our entropy model which addresses the same research topic from an information-theoretic point of view.

Previous research shows evidence for behavioral differences between children and adults in terms of their tendency to learn the probability distributions specific to the input or to move away from the statistical specificity of the input and regularize the input by rule induction (Hudson Kam & Newport, 2005; Hudson Kam & Newport, 2009). Children's tendency to generalize was explained by their incomplete cognitive development (maturational constraints), more specifically by memory constraints (children's overall lower working memory capacity – Cowan, 1997; Gathercole, 1998). For example, Hudson Kam & Newport (2005) found that adults exposed to an input where a determiner occurred only 60% of the times with nouns tended to probability match (or reproduce) the occurrences in their subsequent productions, i.e. produced determiners about 60% of the times, whereas children exposed to the same input tended to impose consistency on the language and produced the determiners more often and highly consistently. The authors suggested that due to their overall encoding limitations and lower working memory capacity, children are more likely than adults to forget the specific statistical properties (probability distributions) of the input, and thus to generalize.

However, adults were also found to regularize the input under specific circumstances (Hudson Kam & Newport, 2009; Hudson Kam & Chang, 2009). Hudson Kam & Newport (2009) found that, when exposed to a *noisier input*, as dubbed by the authors, i.e. 2 main determiners (consistently determining 2 noun classes) and up to 16 *noise* determiners (which could randomly occur with any noun), adults produced the main determiners with nearly 90% of the nouns, indicating they generalized when the input was more complex or *noisier*. Thus, Hudson Kam & Newport (2009) showed that adults can overregularize, just like children, but crucially only when exposed to an overloading *noisier input*, i.e. a highly inconsistent input with many low frequency items occurring randomly, which might have taxed their storing and retrieval capacities enough to cause failures in retrieving the specific forms to probability match their input. These findings are in line with results from studies that employed a different type of generalization task (Reeder et al., 2009; 2013), which showed adults have a higher tendency to generalize *X* as a category in an AXB grammar (rather than just memorize the exact strings), when the input contains increased variability and a heterogeneous distributional structure, in other words a *noisier* input.

Taken together, all these findings of different generalization behavior among young and adult learners were interpreted to bring evidence for the *Less-is-More hypothesis*, namely that children fail to retrieve low-frequency forms and will overuse the more frequent items, while adults having better storing and retrieval capacities are able to produce most of the items in their received input. We reinterpret these findings to be preliminary evidence to an approach that links generalization to a particular type of randomly variable – *noisy* – input and cognitive limitations of the learners. However, the *noisy* circumstances and the extent to which limitations on cognitive capacities account for rule induction are still underspecified. Is better generalization (regularization) helped by limitations on retrieving only or do limitations on encoding also have the same effect?

In a systematic attempt to clarify and better specify *how* and *why* memory limitations could have an impact on generalization, Perfors (2012) further investigated the effect of memory limitations on regularization, but, crucially, during the encoding stage. The author manipulated working memory load during encoding of a simple artificial grammar (noun – determiner pairs) in seven different experimental conditions and found no effect of working memory load on regularization behavior during encoding. In these experiments, participants' generalization behavior was assessed using a word-learning task based on Hudson Kam and Newport (2009), in which participants were presented with 10 bisyllabic non-sense words (called "nouns", since they were simultaneously associated with a visual object) paired with a monosyllabic non-sense word 60% of the times (dubbed "main determiner"), and with other four monosyllabic words, each of them 10% of the times (called "noise determiners"). Besides the target material to be learned (i.e. the "noun-determiner" pairs), there was also non-target (stimulus-irrelevant) material added in the experiment, to obtain different levels of cognitive load. Namely, there was a (control) No-load condition, where participants performed the word-learning task as described, and six Load conditions, in which the word learning task was either interleaved with another verbal or operational task, or simultaneously carried out with a verbal or operational task, taxing the working memory either with a low or a high load. There was no evidence of an effect of working memory load on regularization in none of these conditions. Perfors (2012) also measured participants' working memory in a complex span task (such tasks are widely used to measure working memory capacity – Conway et al., 2005; Unsworth, Spillers, & Brewer, 2009) and found that individual differences in working memory did not predict the generalization behavior.

Other studies investigated the domain-generality of the *Less-is-More hypothesis*, and found that cognitive constraints are reflected in the regularization behavior in non-linguistic domains (Kareev, Lieberman, and Lev, 1997; Ferdinand, Kirby, and Smith, 2019), while other studies found that regularization tendencies and patterns are very similar across domains and language levels – morphology vs word order (Saldana, Smith, Kirby, & Culbertson, 2017).

Summarizing previous findings, qualitatively different types of generalization (*item-bound* and *category-based generalization*) have been investigated in different types of tasks (word-learning, category-based grammars like AXB, object-naming, etc.) and under several types of cognitive load: either increasing the complexity of the target material (i.e. for regularization), or by adding non-target material in concurrent load tasks. These different concurrent tasks taxed the different cognitive capacities – working memory, attention or sub-components of these capacities – in different ways, either sequentially or simultaneously. Overall some of these findings show evidence for a *gradient of generalization* which results from *input variability*, and also from a specific kind of cognitive load under specific *noisy environments* in adults. On the other side, other ways of taxing the relevant cognitive capacities at the time of encoding in noisy environments did not yield the same outcome. So, what is going on here?

These studies show that, while there is some evidence for the *Less-is-More* hypothesis on memory constraints modulating rule induction, it is not yet clear under *what* specific circumstances and *why* memory constraints should have a certain effect on rule learning. Based on previous work briefly reviewed above, two main factors were proposed to be a driving force in rule induction: probability distribution of the exposure language (input entropy) and cognitive constraints on the learning process. However, it is still underspecified *what* pattern of input variability (i.e. what kind of *noisy input*) and *what* specific cognitive load drive rule induction, since findings seem to be conflicting (in some studies or experimental conditions these factors lead to better generalization – Hudson Kam & Newport (2009), in others not – (Hudson Kam & Newport, 2005; Perfors, 2012).

In order to answer these questions, here we employ and further extend the entropy model that we proposed in Radulescu et al. (2019; 2020). This model offers an extended and more refined information-theoretic approach to the *Less-is-More hypothesis*, by bringing together both factors (input entropy and cognitive capacity) in one formula. In our information-theoretic approach, we take a step further from the algorithmic level (i.e. cognitive limitations of the memory and attentional resources) to the computational level (as per terminology by Marr (1982)), i.e. *channel capacity* – our time-dependent noisy information processor. While in Radulescu et al. (2021) we looked at the *time variable* of the *channel capacity*, in this study, we focus on the effect of the *noise variable* which renders a *noisy-channel capacity*, in information-theoretic terms (Shannon, 1984; Radulescu et al., 2021). Specifically, we propose that the effect of *noisy-channel capacity,* in information-theoretic terms, should be disentangled from the effect of overloading the relevant cognitive capacities with additional concurrent tasks. In this proposal, we hypothesize that *noise* (i.e. random stimulus-irrelevant material) adds sufficient entropy in the environment in order to drive a change in the encoding mechanism to move from an *item-bound generalization* to a *category-based generalization.* In the following sections, we first briefly introduce the entropy model supported by our previous findings, then we focus on defining the *noise* variable of the *channel capacity*. Next we formulate the hypotheses of the model about the effect of the *noisy-channel capacity* on rule induction, and we specifically disentangle this hypothesis in information-theoretic terms from previous hypotheses regarding the effect of overloading the limited cognitive capacities assumed to be involved in rule induction. In the remaining sections, we present an artificial grammar learning experiment based on the lowest entropy version of the XXY grammar employed in our previous studies (Radulescu et al., 2019; 2020). In this experiment, while exposing adults to this grammar, we played random digits and beeps in the background, in order to create a *noisy environment,* that would render a *noisy-channel capacity.* Although the *input entropy* was low, we found higher tendency towards *category-based generalization* when there was noise in the background, but crucially only when no simultaneous operational task was required on the noise material.

2. An entropy model for rule induction

2.1 Brief introduction to the model and previous findings

In Radulescu et al. (2019), we proposed an information-theoretic model with the main hypothesis that rule induction is driven by a single mechanism, as a consequence of the dynamics between input properties and the design of the encoding system. More specifically, rule induction, with its two flavors – *item-bound* and *category-based generalizations* – is an encoding mechanism resulting from the interaction between two main factors: *input entropy* and the *channel capacity* of the encoding system (i.e. the amount of entropy that can be encoded per unit of time).

In our model, we use the concepts and formulas for entropy and channel capacity as they were introduced and mathematically demonstrated by Shannon (1948), and we propose a method of calculating input entropy of an XXY artificial language (Radulescu et al., 2019) and a method to estimate the maximum rate of information transmission, i.e. *channel capacity,* of the learner of such an artificial language (Radulescu et al., 2021).

In Radulescu et al. (2019), we exposed adults to a 3-syllable XXY artificial grammar, where we designed six experimental conditions with different input entropy – 2.8, 3.5, 4, 4.2, 4.58, 4.8 bits – which we calculated using Shannon's entropy formula. For a random variable *X*, with *n* values $\{x_1, x_2 \ldots x_n\}$, Shannon's entropy (Shannon, 1948), denoted by *H(X)*, is defined as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) log p(x_i)$$ [29];

where $p(x_i)$ is the occurrence probability of x_i. This quantity (H) measures the information per symbol produced by a source of input, relative to all the possible symbols (values) contained by the set (Shannon, 1948). Results showed that when input entropy increases, the tendency to move from *item-bound* to *category-based generalization* increases gradually (Radulescu et al., 2019).

While previous studies used several entropy measures in order to investigate regularization patterns (Ferdinand, 2015; Ferdinand et al., 2019; Perfors, 2012; Perfors, 2016; Saldana et al., 2017; Samara, Smith, Brown, and Wonnacott, 2017), our model takes a step further and proposes a quantifiable information-theoretic approach that captures not only the effect of entropy on generalization, but the dynamics between the *input entropy* and the relevant encoding capacity (i.e. *channel capacity*). Thus, in order to address this specific dynamics, in Radulescu et al. (2021), we further tested our entropy model by probing the effect of *channel capacity* on rule induction. According to Shannon (1948), *channel capacity (C)* determines the maximum amount of entropy that can be transmitted reliably through a communication channel per unit of time, by using an adequate encoding method. Hence, our entropy model hypothesizes that our finite encoding capacity, i.e. *channel capacity,* places an upper bound on the amount of entropy that can be encoded per unit of time by using an adequate encoding method. An amount of entropy higher than *channel capacity* supports, drives the transition from an encoding method – *item-bound generalization* – to another encoding method – *category-based generalization* – which is more adequate to encode higher entropy. Thus, we probed the effect of the time variable of *channel capacity*. Specifically, we sped up the source rate of information transmission (H'), that is the average amount of entropy (bits) produced by the XXY grammar per second, in order to attempt to exceed the *channel capacity* of the learners who were being familiarized with the grammar. Results showed an increase in the tendency towards *category-based generalization* when the inflow of information per second was higher, even though the statistical properties of the language showed low *input entropy*.

2.2 The noise variable of the channel capacity

In this study, we further develop the concept of *noisy-channel capacity,* in order to extend our model to capture rule induction as an encoding mechanism that develops in accord with Dynamic Systems Theories, as a natural automatic means of adapting to *noisy* (= increasingly entropic) environments. Specifically, we further develop and extend the entropy model by investigating the *noise variable* of the *channel capacity*.

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

Since information transmission in a noiseless environment is nearly impossible to obtain in real life conditions, when defining the *channel capacity,* Shannon (1948) took into account the fact that information transmission happens in noisy environments. In short, Shannon describes the process of information transmission as follows: an information source produces a message, which is encoded by a transmitter into a signal (= the sent signal) to be transmitted to a destination. The encoding method needs to be adequate and efficient for the transmission of the information through a medium, i.e. a channel of transmission. The source transmits the signal at a certain rate per unit of time – *source rate of information transmission*, i.e. *input entropy per second (H')*. The signal reaches a receiver, which performs the decoding operation on the received signal in order to reconstruct the message to deliver to the destination. In time, the noise perturbs the signal, such that the received signal does not match the signal sent out by the source. The received signal is actually a function of the transmitted *signal* (S) and the *noise* (N), i.e. *f(S, N)*. The information content of the signal and the noise, as well as that of the noise-affected received signal can be quantified using *entropy (H)*. In the ideal noise-free case, the amount of information sent by the source equals the amount of received information, i.e. the entropy of the source signal equals the entropy of the received signal. However, if the transmission medium is noisy, for example, a *noisy channel,* during transmission the noise introduces errors, which leads to a loss in information, i.e. missing bits of information. As a result, the received signal does not always match the sent signal, and thus there is *uncertainty* when decoding the sent signal and reconstructing the message. This uncertainty is defined as rate of equivocation (E) (Shannon, 1948), and it basically quantifies the missing bits of information in the received message due to a noisy transmission. It must be specified that the process of information transmission encompasses all processes of information transmission from the source to the destination, that is all the transmission and encoding – decoding processes.

Shannon (1948) argued and demonstrated that the *noisy-channel capacity* (C) is the maximum actual rate of transmission (R) of information, which can be obtained, but crucially only if the encoding method is adequate and efficient:

$$C = \mathrm{Max}\ (R) = \mathrm{Max}\ (H' - E),$$

where H' is the source rate of information transmission, and E is the rate of equivocation.

In simple words, the maximum rate of transmission, i.e. the *channel capacity,* can be achieved by employing an adequate and efficient method of encoding, such that the rate of equivocation (E) is kept at a minimum, so that the actual rate of information transmission is as close as possible to the source rate of transmission. That means that the received signal will be as close as possible to the sent signal. One aspect needs clarification here, namely that the *actual rate of information transmission* (R) is different from the *source rate of information transmission* (H'), as it takes into account the loss of information due to noise (E), which happens in the course of transmission of the information from the source to the destination. The *source rate of information transmission* (H') is the rate at which the source produces and transmits information (i.e. the source rate of information production), while the *actual rate of information transmission* (R) is quantified at the other terminal end, i.e. the receiver, after the noise had caused a loss in information (E).

According to Theorem 11 of the noisy channel transmission (Shannon, 1948), given a certain source with a rate of information production (H'), when H' is less than C, information can be sent through a noisy channel at the rate C (i.e. *channel capacity*), with a very small rate of equivocation (E), if and only if a proper encoding method is used. If H' is higher than C, it is possible to find an adequate encoding method for the signal, such that the rate of equivocation is minimum, but the rate of transmission can never be higher than C. If there is an attempt to exceed the rate C, by using the same encoding method, then there will be an equivocation rate at least equal to the excess rate of transmission.

It follows that, the efficiency of the encoding method is defined by the ratio of the actual rate of transmission to the capacity of the channel. If the encoding method is maximally efficient, the equivocation rate (E) is minimum, so the actual rate of transmission (R) approaches its maximum, which is the channel capacity: C = Max(H' - E) = Max(R). In the ideal noiseless case (where E = 0), R/C = 1, because R = C. If the encoding method is less than maximally efficient, the equivocation rate is higher than 0 (E > 0), so R is lower than C, thus, R/C < 1. In other words, an encoding method is efficient if the equivocation rate is minimum in order for the rate of transmission to achieve its maximum to match the channel capacity. If the rate of equivocation increases, the rate of transmission decreases, which drives the need for a more efficient encoding method, in order to achieve a better match to the channel capacity. In noisy environments, noise perturbs the transmission of the signal, which increases the rate of equivocation, as described above. Thus, the rate of transmission decreases, which calls for a more efficient encoding method.

2.3 Predictions of the model about the effect of *noisy-channel capacity* on rule induction

After having defined and described the concepts, we can continue by stating and elaborating on the main predictions of the model about the effect of *noisy-channel capacity* on rule induction. We employ *channel capacity* in our study to model the information transmission system, i.e. the maximum finite rate of information transmission of the learning system. Recall: transmission involves the entire process of information transmission from the source to the destination, that is all the transmission and encoding – decoding processes.

As proposed in our previous studies (Radulescu et al., 2019; Radulescu et al., 2021), *item-bound generalization* and *category-based generalization* are outcomes of the same information encoding mechanism, as a *gradual* transition from a high-fidelity form of encoding (*item-bound generalization)* to a high-generality encoding (*category-based generalization)*. This transition is driven by the interaction between *input entropy* and the finite encoding capacity of the learning system, i.e. *channel capacity*. Here we further extend and elaborate on the predictions related to the effect of the finite *channel capacity,* which were proposed and probed in Radulescu et al. (2021), and we also further investigate and add into the formula the effect of the *noisy-channel capacity* on rule induction.

a. As proposed in Radulescu et al. (2021), if the *channel capacity* is higher than or matches the source rate of information transmission (H' – that is the average number of symbols produced by the XXY grammar per second – input entropy per second), then the information (message) can be encoded by using an encoding method which matches the statistical structure of the input, i.e. the probability distribution of the specific items in the input. Thus, the information about specific items and their configuration (i.e. input entropy) can be encoded with a high-fidelity symbol specificity (i.e. probability matching to the input), and can be transmitted through the channel at the channel rate (i.e. entropy per unit of time) and stored by *item-bound generalization*.

b. Conversely, if the *channel capacity* is lower than the source rate of information production (H'), that is an attempt is made to exceed the finite *channel capacity* of the encoding system, it is possible to find a proper method that encodes more information (entropy), but the rate of transmission cannot exceed the available *channel capacity*. As per Theorem 11 (Shannon, 1948): if H'>C, another encoding method can be found to transmit the signal, but the rate of transmission cannot be higher than C. If there is an attempt to transmit information at a higher rate than C, by using the same encoding method, then there will be an equivocation rate at least equal to the excess rate of transmission. Thus, if the inflow of entropy creates an excess source rate of information, which is higher than the available channel (H' > C), the rate of equivocation (E) increases, if the encoding method is not suitable. Since the *channel capacity* cannot be exceeded, this calls for a more efficient encoding method such that the rate of equivocation is minimized in order for the actual rate of transmission (R) to achieve its maximum, to match the *channel capacity*. Specifically, when the source entropy per second is higher than the available *channel capacity*, the high-specificity *item-bound generalization* becomes inefficient and prone to many errors. Therefore, the information cannot be encoded with a high-fidelity method (i.e. probability matching to the input), because this encoding method gives rise to a high rate of equivocation. As explained before, a high rate of equivocation calls for another more efficient method of encoding. Thus, we hypothesize that the excess of entropy entering the channel results into breaking bindings between items, and reorganizing the redundant (shared) and non-redundant (specific) features of discrete symbols in order to erase or "forget" insignificant features. This leads to a compression of the signal by reducing the specific features encoded with individual items and re-grouping them in categories. As a result, a new form of encoding is created, which allows for higher *input entropy* to be encoded at the same *channel capacity*, but by yielding a more general (less specific) *category-based encoding.*

In Radulescu et al. (2021) we hypothesized that it is precisely the *finite channel capacity* that drives restructuring of the information, in order to find another form of encoding, i.e. *category-based generalization,* which is more efficient at maximizing the rate of transmission, that is by minimizing the rate of equivocation. Here we further extend the finite *channel capacity* hypothesis and we formulate a specific hypothesis by emphasizing on the effect of *noise*:

c. As explained in the previous section, the efficiency of the encoding method is defined by the ratio of the actual rate of transmission to the capacity of the channel. Thus, we hypothesize that *noise* adds sufficient entropy per second which enters the channel, in order to drive a change in the encoding mechanism to find a more efficient encoding method. More specifically, noise inflow perturbs the signal and increases the rate of equivocation (as explained above). Since noise increases the rate of equivocation, and an increased rate of equivocation calls for a more efficient encoding method, we expect that an increase in noise should accelerate the drive towards a reorganization of the information, such that a more efficient encoding method is found. This hypothesis is in line with Dynamic Systems Theory (DST), according to which random perturbations in the environment (i.e. noise) add to the input entropy and accelerate self-organization into a new structure (Stephen, Dixon, & Isenhower; 2009).

As explained above, *channel capacity* is used here to model the finite encoding capacity of the learning system in information-theoretic terms (i.e. at the computational level, in the sense of Marr (1982)). In psychological terms (at the algorithmic level), we follow experimental evidence from the *Less-is-More hypothesis* line of research, which suggests that memory constraints drive rule induction (Hudson Kam & Newport, 2005; Hudson Kam & Newport, 2009), and embed this in classical and more recent models of memory and attention (Baddeley, Eysenck, and Anderson, 2015; Cowan, 2005; Oberauer and Hein, 2012). Hence, we hypothesize that the cognitive capacities that underlie *channel capacity,* specifically in linguistic rule induction (and, implicitly, in category formation), are the attentionally-controlled regions of activated long-term memory, in other words working memory (WM). Rule induction can be argued to rely on the storage and online time-dependent processing capacities that support the ability to maintain active goal-relevant information (the rule) while concurrent processing (of other possible hypotheses, and of noise) takes place (which is what defines WM as well – Conway et al., 2002). Corroborating evidence comes from positive correlations found between WM and domain-general categorization tasks (Lewandowsky, 2011).

As we argued in Radulescu et al. (2021) and Chapter 3, we generally deem linguistic rule induction to be supported by a domain-general WM capacity, rather than language-specific algebraic rule learning as proposed by early prominent research (Marcus et al., 1999). In the current study we further explore the effect of one of the components underlying *channel capacity* in linguistic rule induction, namely what we dubbed as a domain-general pattern recognition capacity. The rationale is that a rule induction task can be intuitively envisaged as a task of finding patterns/rules in the input.

As proposed in our previous studies (Radulescu et al., 2021, and Chapter 3) a possible candidate test of domain-general pattern recognition is the Raven's Standard Progressive Matrices (RAVENS test – Raven, Raven, & Court, 2000), which was shown to be based on rule induction (Carpenter, Just & Shell, 1990; Little, Lewandowsky, & Griffiths, 2012) and to rely on similar storage and online time-dependent processing capacities to maintain active goal-relevant information (the rule) while concurrent processing takes place (Conway et al., 2002). Importantly, this pattern recognition test and WM capacity are not identical (Conway et al., 2003), and WM is not a causal factor for pattern recognition (Burgoyne, Hambrick, & Altmann, 2019). However, high positive correlations were found between measures of WM capacity and tests for this domain-general pattern-recognition capacity (like RAVENS – e.g. Conway et al., 2002; Little, Lewandowsky and Craig, 2014; Dehn, 2017).

In the Conclusion section, we will come back to this topic for a discussion of the compatibility between current memory and attention models (at the algorithmic level) and our entropy model (at the computational level).

Development of these hypothesized underlying cognitive capacities entails as an effect a developmental increase in *channel capacity*, which leads to a higher amount of entropy that can be encoded per unit of time. It follows from the previous predictions of the model that a developmental increase in *channel capacity* reduces the need and the tendency to move to a more general *category-based* form of encoding. This prediction of the model would explain the differences in regularization behavior observed between young and adult learners who are exposed to the same input entropy: adults have a lower tendency to encode the input as *category-based encoding* than young learners, because adults' *channel* has a higher information encoding rate.

3. Effect of noisy channel in information-theoretic terms vs. previous hypotheses regarding cognitive constraints on rule induction

Noisy-channel capacity according to our model differs from the previous hypotheses of cognitive constraints on rule induction, as they were proposed and investigated in previous studies mentioned above (e.g. *Less-is-More hypothesis* with its versions and follow-up studies). Next, a specific prediction will be made on the effect of an attempt to exceed the *noisy-channel capacity* on rule induction, as we propose it should be disentangled from the effect of overloaded working memory capacity with additional tasks.

Theoretically, following the above-mentioned definition of *channel capacity* (i.e. the amount of entropy, including *noise*, that can be transmitted per unit of time), and Shannon's Theorem 11, an attempt to exceed the *channel capacity* in an artificial grammar experiment can be obtained in two straightforward ways: either by increasing the amount of entropy that enters the channel, or by speeding up the rate of feeding information (entropy) into the channel. When quantifying the entropy in an artificial grammar learning task, in general there are two main sources of entropy that should be factored in: *input entropy* that is the target-intrinsic (or signal-specific) entropy layers (namely acoustic, prosodic, phonological, morphological, semantic, distributional, etc., entropy of the target signal), and also target-extrinsic (or stimulus-unrelated) entropy (or background *noise*). Thus, there are various possible sources for obtaining a *noisy channel* in an artificial grammar learning environment like the one simulated in the experiments carried out by Radulescu et al. (2019). Specifically, the noise, that is perturbations that interfere with the signal, could stem from individual cognitive capacities (e.g. working memory), from different learning strategies that learners employ in order to cope with these finite cognitive capacities, from general biases regarding language composition and structure, from knowledge regarding the discrete symbols (i.e. particular bigrams/trigrams of syllables), etc., but also from external sources, such as stimulus-irrelevant noise in the background. According to our model, all these sources of noise interfere with the actual signal sent, so that the received signal becomes a function of the sent signal and the noise variable.

Since in the formulation of the specific predictions of our entropy model we disentangled between the effect of the *input entropy* (i.e. target-intrinsic entropy) and that of *channel capacity*, it follows that, practically, there are two methods to try and exceed the *channel capacity*, while keeping constant the target-intrinsic entropy:

1. Increase the source rate of production, to directly modulate the time variable of the channel capacity. This method reduces the time that the same amount of input entropy passes through the channel. We employed it in Radulescu et al. (2021) and found that a sped up rate of information transmission, which was higher than the *channel capacity*, drove *category-based generalization* in the same XXY grammar employed by Radulescu et al. (2019).

2. Add stimulus-unrelated entropy (noise) in the input to modulate the total amount of entropy that enters the channel while keeping the time variable constant, in order to render a noisier channel. In this study we used this method, that is we added stimulus-irrelevant entropy (noise) in the input, in order to attempt and exceed the channel capacity of the learners while they were performing a rule induction task on the same XXY grammar employed by Radulescu et al. (2019).

Background noise entropy was shown in a wide range of studies to facilitate learning, in general. More specifically, in the adult problem-solving paradigm, Stephen et al. (2009) presented adults with a series of gear system problems on a computer screen. One group of participants saw the problems appear in a consistent spatial location, while the other group saw the problems in random locations, i.e. non-target entropy (noise) was added to the task. Although both groups eventually abstracted a short-cut to solve the gear problems, the group exposed to a noisy environment did so the fastest. Stephen et al. (2009) explained these findings as an instance of dynamic systems of cognition where non-target, extraneous entropy (i.e. noise) during learning speeds up early learning by helping new cognitive structure emerge via a shift from one stable state to another.

In a word learning paradigm, Twomey, Ma and Westermann (2018) exposed 2-year-olds in a referent selection task to either objects on a uniform white background, or on differently colored backgrounds. At test, only children in the variable background condition showed evidence of learning label-object associations. Authors suggested that these findings fit the dynamic systems theory, which suggests that extraneous entropy (here in the form of background variability) adds sufficient noise to the system to cause a change in behavior which supports learning.

Additionally, in a computational model, which simulates word learning from multiple intrinsic and extrinsic cues, Monaghan (2017) showed that noise in the environment, i.e. less than perfect reliability of such cues or source of information is noisy, supports robustness of learning, even though there is a trade-off with speed of initial learning. Similar beneficial effects of noise on learning were also found in studies on generalization of learned information to a new context (Gartman & Johnson, 1972; Godden & Baddeley, 1980), and in face recognition studies (Smith & Handy, 2014).

In psychological terms (at the algorithmic level), as mentioned above, following suggestions from classical and more recent models of memory and attention (Baddeley, Eysenck, and Anderson, 2015; Cowan, 2005; Oberauer and Hein, 2012), we hypothesize that the underlying cognitive mechanisms that modulate channel capacity are the attentionally-controlled regions of activated long-term memory, i.e. the working memory. In this study, we did not only measure individual differences in cognitive capacities that were hypothesized to underlie the individual channel capacity of the participants, but we also looked at the effect of channel capacity on rule induction from the following perspectives:

1. Firstly, by taxing the relevant cognitive capacities in real time, at the moment of the actual process of information encoding.

2. Secondly, by teasing apart the effect of taxing the actual cognitive capacity with additional tasks (i.e. taxing what we shall dub the "operational

processor", at the computation level), and the effect of an attempt to exceed *channel capacity* (i.e., the actual rate of information transmission), in purely information-theoretic terms, in real time, by introducing additional signal-irrelevant entropy (noise) per unit of time, to achieve a *noisier channel*.

Previous hypotheses regarding cognitive constraints did not clearly specify whether such constraints should be thought of in terms of overloading the capacity with secondary tasks, or in terms of the amount of information fed into the relevant operating capacities. Therefore, previous studies regarding cognitive constraints on rule learning tested these hypotheses either by adding more information (i.e. more target-intrinsic variability) in the learning material (Hudson Kam & Newport, 2009), or by overloading the processing capacities with additional tasks at the time of learning (Perfors, 2012). Therefore, mixed results and conclusions were obtained, as discussed in the introduction of this paper, and as a consequence the hypothesis of cognitive constraints on rule learning remains largely underspecified.

We propose that a clear distinction should be made between the "operational processor", which carries out a certain number of tasks, and the inflow of information fed into it for the purpose of carrying out those tasks, i.e. the rate of information transmission, which is determined by the channel capacity. More specifically, the "operational processor" could be thought of as the central executive component in Baddeley's model – Baddeley et al. (2015) or as one of the functions of the attentionally-controlled working memory (Cowan, 2005; Oberauer and Hein, 2012). The input of information to this task-handler ("operational processor") is determined by the available channel capacity.

As presented in the previous section, our entropy model poses the hypothesis that an attempt at exceeding the channel capacity results in the need to find a more efficient encoding method, i.e. that enables a higher rate of information transmission with less equivocation. The outcome of the need for a more efficient encoding is a higher tendency towards *category-based generalization*. Therefore, in information-theoretic terms, our model poses a specific prediction: the transition to a *category-based encoding* method is driven by the attempt at exceeding the channel capacity with additional signal-irrelevant entropy (noise) per unit of time, to achieve a noisier channel. Indeed, also at the algorithmic level, there is evidence that random material added in the background of learning tasks leaves a traceable footprint on the memory of the participants (Conway, Cowan, & Bunting, 2001; Cowan, Nugent, Elliott, Ponomarev, & Saults, 1999; Cowan, Nugent, Elliott, & Saults, 2000; Oberauer & Lewandowsky, 2016).

To conclude this section, *channel capacity* quantifies the actual rate of information transmission in noisy environments, namely by taking into account the rate of equivocation due to the effect of noise on the sent signal. If an attempt is made to exceed the finite *channel capacity* by increasing the noise in the channel, the encoding method changes into a more abstract encoding in order to allow a higher rate of information transmission with a lower rate of equivocation. Thus, in line with dynamic systems theory deemed to explain the generalization behavior in problem-solving tasks (e.g. Stephen et al.; 2009), our model hypothesizes that the *finite channel capacity* is designed to drive re-structuring of the information in order to shape the encoding into a more general abstract form of encoding, for the purpose of adapting to noisier (=increasingly entropic) environments. In the next section, we propose an experiment to test this prediction, while teasing it apart from the effect of overloading the "operational processor" with additional tasks.

4. Experiment: design and rationale

The goal of this study is to probe the effect of the second main factor of the entropy model, i.e. *channel capacity*, by adding stimulus-irrelevant entropy (i.e. noise) in the background while participants are exposed to the same XXY grammar from Radulescu et al. (2019). The added noise is hypothesized to inject sufficient extraneous entropy in order to drive a change in encoding method, thus yielding a higher tendency towards *category-based encoding.*

We tested the effect of *channel capacity* on rule induction by adding noise to render a noisier channel, while disentangling this effect from the effect of overloading the underlying cognitive capacities with an additional task in real time. To this end, we designed and ran an experiment, in which we employed the lowest entropy version (i.e. 2.8 bits) of the XXY grammar we used in Radulescu et al. (2019), for which we found no evidence of *category-based encoding,* thus we shall use it as the control experiment. In the two experimental conditions of this study, we attempted to exceed the *channel capacity* in the following way: we exposed adults to the XXY language (i.e. the signal), while playing a stream of digits simultaneously (i.e. the noise, or the unpredictable material which was irrelevant to the XXY language learning). Participants in one experimental condition were given two tasks: one task to be performed on the signal (i.e. on the XXY language) and an additional simultaneous task on the stream of digits, that is to perform a memorization operation on the stimulus-irrelevant material. In the other condition, the same noise was played in the background, but crucially there was no additional task required on the noise. The rationale of this design is to tease apart the effect of simply injecting noise in the environment to render a noisier channel from the effect of overloading the "operational processor" with another simultaneous task.

More specifically, participants were exposed (aurally) to an artificial XXY grammar using the same stimuli as those used in the lowest entropy condition in Radulescu et al. (2019) – 2.8 bits. While presenting them with the artificial grammar (the signal), a stream of digits and random beeps (the noise) played in the background. In one condition (Dual-Task Condition), the participants were asked to remember only the digits played right before every beep in the stream and report them in the same order at the end of the familiarization phase. Participants in the other condition (Distractor Condition) were exposed to the exact same signal and noise streams, but crucially they were not assigned any task on the noise stream. Next, just as in the design by Radulescu et al. (2019), in the test phase participants in both conditions were presented with the same grammaticality judgement task, where they had to answer a yes/no question to indicate whether the test strings could be possible in the familiarization language. There were four types of test strings designed to probe how the participants encoded the familiarization stimuli, as presented below.

Familiar-syllable XXY (XXY structure with familiar X-syllables and Y-syllables) – correct answer: yes – accept. This type of test strings was used to test learning of the familiar strings. According to the hypotheses of the entropy model, as the noise was hypothesized to pose an excess on the *channel capacity,* the Distractor group was expected to accept this type of strings as grammatical, either by *item-bound generalization* (i.e. *same-same-different* structure with familiar syllables), or by *category-based generalization* (i.e. *same-same-different* structure with any syllables). The Dual-Task group was also expected to accept these strings as grammatical either by *item-bound* or by *category-based generalization* driven by the background noise. Based on evidence from previous studies with dual tasks (Cocchini, Logie, Della Sala, MacPherson, & Baddeley, 2002; Morey & Mall, 2012; Perfors, 2012; Saults & Cowan, 2007), the Dual-Task group is expected to have an overall worse performance, due to the task-specific challenges, in our terminology, that is the "operational processor" is overloaded with two different tasks, and therefore less storing and processing resources are available since they have to be split between two different tasks.

New-syllable XXY (XXY structure with new X-syllables and Y-syllables) – correct answer: yes – accept. This type was used to test whether learners' *item-bound generalization* was shaped into *category-based generalization* which enables them to accept XXY strings with new syllables (i.e. *same-same-different* structure with regardless of familiar or new syllables). According to the hypotheses of the entropy model, both groups were expected to accept this type of strings as grammatical, as the noise was hypothesized to pose an excess on their *channel capacity.* However, based on evidence from previous studies with dual tasks, again the Dual-Task group is expected to have an overall poorer performance than the Distractor group due to an overtaxed "operational processor". However, absolute mean acceptance rate of this type of strings does not represent direct evidence for *category-based generalization*. As we argued in Radulescu et al. (2019), this mean should be compared to the mean acceptance rate of Familiar-syllable XXY strings: the smaller the difference of the mean acceptance rate (i.e. the effect size) between New-syllable XXY strings and Familiar-syllable XXY strings is, the more likely it is that learners have formed *category-based generalization*.

Familiar-syllable X_1X_2Y (X_1X_2Y structure with familiar syllables) – correct answer: no – reject. According to the entropy model, participants are expected to confidently reject this type of strings, either by having encoded the input as *item-bound* or *category-based generalizations*. Specifically, participants in the Distractor condition are expected to confidently reject this type of strings, as their tendency towards *category-based generalization* is hypothesized to be driven by the effect of the noisy channel. However, based on previous evidence from the dual-task literature, participants in the Dual-Task condition are expected to have difficulties to confidently reject the Familiar-syllable X_1X_2Y strings as deviant from the XXY language, since their "operational processor" is overloaded with a simultaneous memorization task, which impairs the formation of *item-bound generalization* (i.e. *same-same-different* structure with familiar syllables).

New-syllable X_1X_2Y (X_1X_2Y structure with new syllables)) – correct answer: no – reject. In both conditions, participants are expected to confidently reject this type of strings, either by having encoded the input as *item-bound* or *category-based generalizations*.

In addition to probing the direct effect of the attempt to exceed *channel capacity* in real time, as presented above, we also measured the individual differences in relevant cognitive capacities on rule induction: memory capacity and a domain-general pattern-recognition capacity. To this end, we tested each participant on three independent tests: a Forward Digit Span task, which is a measure of explicit short-term memory (Baddeley et al., 2015), an incidental memorization task, which measures implicit memory capacity (Baddeley et al., 2015), and RAVENS Standard Progressive Matrices (Raven et al., 2000), which is a standardized test based on visual pattern-recognition (Carpenter et al. 1990, Little et al. 2014). According to the hypotheses of our entropy model, we predicted a positive effect of RAVENS on the tendency to move from an *item-bound* to a *category-based generalization*, and a negative effect of the explicit/incidental memory tests on the same transition from one type of encoding to the other.

To the best of our knowledge this is the first artificial grammar experiment that investigates the effect of *noisy-channel capacity* in rule induction, by specifically testing information-theoretic predictions made by the entropy model in order to disentangle the effect of *noisy-channel capacity* from the effect of overloading the underlying cognitive capacities with additional tasks.

5. Methods

5.1 Participants

60 healthy, non-dyslexic Dutch speaking participants (42 females, 18 males, age 19-42, M=22.75) were randomly assigned to either the Dual-Task condition or the Distractor condition. Only healthy participants that had no known language, reading or hearing impairment or attention deficit were included. They all signed a form of consent and were paid for their participation.

5.2 Tasks and materials

Task 1: XXY grammar

Familiarization stimuli. In both the Dual-Task and the Distractor conditions, participants listened to the same XXY artificial grammar used in the low entropy condition of Experiment 2 from Radulescu et al. (2019). The three-syllable strings of the language display an XXY structure (each letter stands for a set of syllables), namely each string has two identical syllables (XX) followed by another different syllable (Y): e.g. keːkeːmy, daːdaːli . All syllables are natural Dutch syllables having the same structure, i.e. a consonant followed by a long vowel. Seven X-syllables and seven Y-syllables (the subset of X-syllables does not overlap with the subset of Y-syllables) were used to generate seven strings (see Appendix A for complete stimulus set). All seven strings were repeated four times (7 strings * 4 = 28 strings) in each familiarization phase (there were three familiarization phases, each consisted of the same 28 strings). Thus, the entropy was the same in all familiarization phases – 2.8 bits. We employed the same method for the entropy calculations as in Radulescu et al. (2019), which is a fine-tuned extension of a related entropy calculation method proposed by Pothos (2010) for finite state grammars (see Table 1 below for complete entropy calculations). The order of presentation of the strings was randomized for every participant.

The stream of digits and beeps. For each familiarization phase, we created a stream of digits interleaved with random beeps (i.e., three such audio streams). For example, 8-2-3-BEEP -6-5-7-8-BEEP -9-5-8-BEEP -6-9-2-4-BEEP -7-3-6-7-0-BEEP -6.

Low Entropy
H[bX]=H[7] = -Σ[0.143*log0.143] = 2.8
H[XX] = H[7]= 2.8
H[XY] = H[7] = 2.8
H[Ye] = H[7] = 2.8
H[bXX] = H[7] = 2.8
H[XXY] = H[XYe]= H[7] = 2.8
H[bigram] = 2.8
H[trigram] = 2.8
H[total] = $\frac{\mathbf{H[bigram]+H[trigram]}}{\mathbf{2}}$ **=** 2.8

Table 1. Entropy value. Taken from Radulescu et al. (2019)

Test stimuli. The three familiarization phases were interleaved with three (quick) intermediate test phases and a final (longer) test phase. Each intermediate test phase included four test strings, one of each of the four types presented in the previous section. The final test had eight test strings (two of each type). Thus, in total, there were (4+4+4+8=) 20 test strings (see Appendix B for the complete set of stimuli). Accuracy score for the learning of the XXY grammar was measured as correct acceptance of Familiar-syllable XXY and New-syllable XXY strings, and correct rejection of Familiar-syllable X_1X_2Y and New-syllable X_1X_2Y strings.

Task 2: Forward Digit Span

Participants were explicitly told that this was a memory test, during which a series of digits would be presented aurally, and they would have to recall them in the same order. To prevent participants from creating a visual pattern on the keypad while listening to the digits, we modified the standard Forward Digit Span task such that no physical keyboard was made available to the participants, rather a row with buttons for each digit was displayed in a line on the screen only in the moment when they were asked to enter the digits by clicking the buttons, and disappeared during the listening phases. We used the standard scoring method: we measured the highest span of each participant and recorded it as one data point per participant.

Task 3: Incidental Memorization Test

Participants listened to 30 bisyllabic nonsense words resembling Dutch phonology and phonotactics. Crucially, participants were not told in advance that a memory test would be administered. They were only told that they were about to listen to words from a forgotten language. They were instructed to imagine what the word might have meant in the forgotten language and to pick a category (flower, animal, or tool) for each word they heard, based on what the word sounded like to them. They had 3 seconds to choose a category for each word, by pressing the button for flowers, animals, or tools.

After this phase was over, a surprise message appeared on the screen, informing the participants that they would be given a memory test, which would check whether they remembered the words they categorized during the previous phase. They were instructed to indicate whether they heard the word previously, by clicking a yes/no button on the screen. The memorization test consisted of 13 targets and 13 foils.

Task 4: RAVENS

Participants had to solve 5 sets of matrices, with 12 matrices per set. Each matrix consists of a nine visual patterns arranged in a particular order in accordance with some underlying rules, of which one pattern is missing. Participants have to solve the matrices by finding the missing pattern in a multiple-choice task.

5.3 Procedure

Participants were tested in a sound-proof booth, and they completed the tasks in the order presented above.

For Task 1 – XXY grammar, the participants were instructed to listen a "forgotten language" that would not resemble any language that they might be familiar with, but which had its own rules and grammar. The instructions informed participants that the language had more words than the ones played in the familiarization phases. They were also told that there would be three familiarization phases interleaved with three intermediate tests and a final (longer) test phase. Participants were explained that the tests were meant to check what they had noticed about the language. They were asked to judge, by pressing a Yes/No button, whether the test words could be possible in the language that they listened to. Participants were also told that another audio stream would play simultaneously with the "forgotten language", and specifically a stream of digits interleaved with beeps. Crucially, here the instructions were different for the two experimental groups, as follows. Participants assigned to the Dual-Task Condition were instructed to remember all the digits right before every beep, and report these digits, in the same order, at the end of each familiarization phase. For the example stream presented above (8-2-**3**-BEEP -6-5-7-**8**-BEEP -9-5-**8**-BEEP -6-9-2-**4**-BEEP -7-3-6-7-**0**-BEEP -6), they would have to report: 38840. Participants were instructed that immediately after each familiarization phase they would have to report the digits (by manually entering them in a field on the computer screen). They were also warned in advance that in each phase a different stream of digits would be played, so after reporting each set of digits they could forget it. Participants assigned to the Distractor condition were only informed that they would hear a stream of digits and beeps simultaneously with the "forgotten language", but crucially they were not given any task to perform on them. In order to match the design in terms of time and requirements, participants in the Distractor condition were asked to enter a random set of five digits after each familiarization phase. After entering the set of digits, either remembered from the simultaneous stream or random, both groups of participants would go ahead with each intermediate test, and then continue with the next familiarization phase. This would continue until the end of the task, which lasted around 5 minutes.

Next, they were given the instructions for the Forward Digit Span, namely they were explicitly instructed that it was a memory test where they would listen to streams of digits, which they would have to recall in the same order. This task lasted around 5 minutes.

The third task was the Incidental Memorization task, for which they were instructed to listen to the words from another "forgotten language" and to imagine what their meaning was, based on how the words sounded like. Importantly, they were not told in advance that a memory test would follow. This task lasted about 7 minutes.

Next, participants were asked to perform the RAVENS matrices test, which was a paper-and-pen task that they had to solve while seated at the desk.

The standard RAVENS task allows participants to spend 50 minutes in total, but, after running a pilot testing, we modified the task to allow participants only a shorter amount of time (35 minutes) to complete the task, in order to avoid an overall time-consuming and exhausting experimental session. The experimenter would walk in 20 and 30 minutes after participants started the session, to announce the remaining time. The entire experiment lasted about one hour.

5.4 Data scoring and analysis

For Task 1, we recorded all the yes/no answers and coded them as correct or incorrect according to the criteria presented for each type of strings in Section 4 above. From all the 20 correct/incorrect answers for each participant we calculated a proportion of correct answers per each type of test item. Next, instead of directly analyzing proportions we performed an empirical logarithmic transformation, in order to analyze the data using a linear regression model.
In the Forward Digit Span task, we used the standard scoring method, that is we measured the highest span of each participant and recorded it as one data point per participant. In the Incidental Memorization Task, we recoded all correct/incorrect answers into hits and false alarms, and we calculated a *d'* value for each participant. For the RAVENS test, we used the standard scoring method, that is we counted all correct answers to all sets of questions, and then we used the standard RAVENS tables to transform them into age-corrected percentiles.

6. Results

Figure 2 shows the mean accuracy rate (proportion of correct acceptance of the grammatical test items and correct rejection of the ungrammatical ones) for all test string types, across the two experimental conditions (Groups) – Dual-Task and Distractor. The mean correct acceptance rate for Familiar-syllable XXY strings is *M* = .87 (*SD* = .33) in the Dual-Task group, while in the Distractor group it is *M* = .94 (*SD* = .23). The mean correct rejection rate for Familiar-syllable X1X2Y strings in the Dual-Task group is *M* = .59 (*SD* = .49), while in the Distractor group it is *M* = .89 (*SD* = .31). The mean correct acceptance rate for New-syllable XXY strings is *M* = .59 (*SD* = .49) in the Dual-Task group, while in the Distractor group it is *M* = .75 (*SD* = .43). The mean correct rejection rate for New-syllable X1X2Y strings in the Dual-Task group is *M* = .81 (*SD* = .39), while in the Distractor group it is *M* = .95 (*SD* = .22).

Figure 3 shows the distribution of individual mean rates per test type in each experimental condition, Dual-Task and Distractor.

In order to probe the effect of *noisy-channel capacity* on rule induction, we compared the performance in the two conditions (Dual-Task and Distractor groups) in a general linear mixed effects analysis of the relationship between Accuracy (correct acceptance of the grammatical test items and correct rejection of the ungrammatical ones) and Type of Test (Familiar-syllable XXY, New-Syllable XXY, Familiar-syllable X1X2Y, New-Syllable X1X2Y), Group (Dual-Task, Distractor), as well as Group x Type of Test interaction. Therefore, as dependent variable we entered Accuracy score into the model. As fixed effects we entered Type of Test, Group and Group x Type of Test interaction. The scores for Forward Digit Span, Incidental Memorization Task and RAVENS tests were entered one by one as covariates in the model. As random effect we had an intercept for subjects. 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. The model reported here is the best fitting model, both in terms of the model's accuracy in predicting the observed data, and in terms of AIC (Akaike Information Criterion).

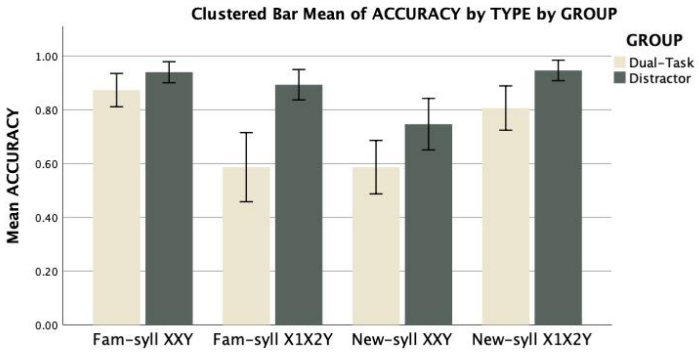


Figure 2. Proportions of correct acceptance for XXY (Familiar- and New-Syllable) strings and proportions of correct rejection for X1X2Y (Familiar- and New-Syllable) strings. Error bars show standard error of the mean.

Error Bars: 95% CI

Error Bars: +/- 2 SE

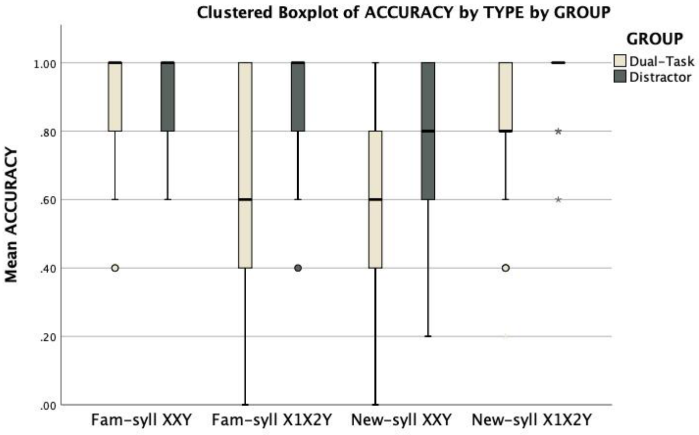


Figure 3. The distribution of individual mean accuracy rates per test type in each condition, Dual-Task and Distractor

We found a significant main effect of Type of test strings (F(3, 180.000) = 13.910, *p* < .001), a significant Group x Type interaction (F(4, 119.318) = 5.542, *p* < .001), a significant Group x Incidental Memorization Task interaction (F(2, 60) = 3.035, *p* = .05), and a non-significant Group x RAVENS interaction (F(2, 60) = .275, *p* = .76).[30]

Pairwise comparisons of the Estimated Marginal Means (adjusted to the mean values of the covariates in the model, i.e. Incidental Memorization Task = 1.547, RAVENS = 67.5) revealed a significant difference between Groups (Dual-Task and Distractor groups) for the Familiar-syllable X1X2Y (M = .35, SE = .067, F(1, 185.452) = 27.178, *p* < .001) and the New-syllable XXY (M = .14, SE = .067, F(1, 185.452) = 4.592, *p* = .033). For the other two Types of test, pairwise comparisons of the Estimated Marginal Means adjusted for the same level of the covariates revealed a non-significant difference between Groups (Dual-Task and Distractor group): Familiar-syllable XXY (M = .026, SE = .067, F(1, 185.452) = .147, *p* = .702) and New-syllable X1X2Y (M = .077, SE = .067, F(1, 185.452) = 1.300, *p* = .256).

The Incidental Memorization Task had a significant positive effect on the overall Accuracy scores (across Test Type strings) in the Dual-Task group (*t(60)* = 2.452, *p* = .017, R^2 = 0.04), but a non-significant effect on the overall accuracy scores in the Distractor group *t(60)* = -.238, *p* = .81.

Further, Cohen's effect size value for the mean difference in correct answers between the Dual-Task and the Distractor groups was d = .24 (Familiar-syllable XXY), d = .73 (Familiar-syllable X1X2Y), d = .34 (New-syllable XXY) and d = .44 (New-syllable X1X2Y). The effect size for the difference between acceptance of Familiar-syllable XXY vs. New-syllable XXY was higher in the Dual-Task group (Diff of Means = 0.28, *d* = 0.67) compared to the same difference in the Distractor group (Diff of Means = 0.19, *d* = 0.55).

7. Comparing this experiment with the single-task noiseless experiment from Radulescu et al. (2019)

Although the present experiment was carried out at a later stage than the experiments reported in Radulescu et al. (2019), and employed additional tasks to control for the individual differences of participants (i.e. Incidental Memorization Task, RAVENS), and the number of participants was double compared to the experiments in Radulescu et al. (2019), we think that a comparison with the lowest entropy condition (2.8 bits) from that study would be in order. The reason is that the first task of this experiment was basically a follow-up task based on the exact artificial grammar, stimuli and procedure from that study, with only an additional stream of noise played in the background, and an additional task in the Dual-Task condition only. Thus, we compared the Dual-Task condition with the lowest entropy condition from Radulescu et al. (2019),

[30] None of the other factors or covariates had a significant effect, and since they did not improve the model they were removed from the final model reported here.

which we will name here the Single-Task condition, and we also compared the Distractor condition with the Single-Task condition.

For comparison reasons and for convenience, we will briefly present here the descriptive statistics for the Single-Task condition: Familiar-syllable XXY (M = .95, SD = .22), New-Syllable XXY (M = .57, SD = .5), Familiar-syllable X1X2Y (M = .83, SD = .37), New-Syllable X1X2Y (M = .92, SD = .27) (Radulescu et al., 2019). Figure 4 shows the mean accuracy rate (proportion of correct acceptance of the grammatical test items and correct rejection of the ungrammatical ones) for all test string types, across the three experimental conditions (Groups) – Dual-Task, Distractor and Single-Task.

In order to probe the effect of *an additional overloading task* on rule induction, we compared the performance in the two conditions (Dual-Task and Single-Task) in a general linear mixed effects analysis of the relationship between Accuracy (correct acceptance of the grammatical test items and correct rejection of the ungrammatical ones) and Type of Test (Familiar-syllable XXY, New-Syllable XXY, Familiar-syllable X1X2Y, New-Syllable X1X2Y), Group (Dual-Task, Single-Task), as well as Group x Type of Test interaction. Therefore, as dependent variable we entered Accuracy score into the model. As fixed effects we entered Type of Test, Group and Type of Test x Group interaction. As random effects we had intercepts for subjects and items. 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. The model reported here is the best fitting model, both in terms of the model's accuracy in predicting the observed data, and in terms of AIC (Akaike Information Criterion).

We found a significant main effect of Type of test strings (F(3, 126.000) = 8.890, *p* < .001), a non-significant main effect of Group (F(1, 42) = 1.034, *p* = .315), and a non-significant Type x Group interaction (F(3, 126.000) = .736, *p* = .533).

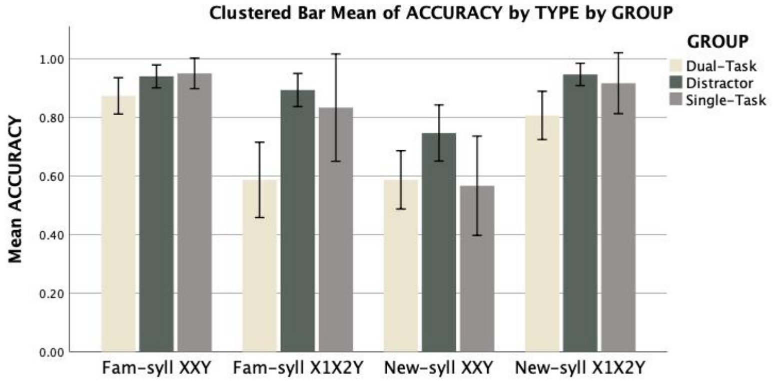


Figure 4. Proportions of correct acceptance for XXY (Familiar- and New-Syllable) strings and proportions of correct rejection for X1X2Y (Familiar- and New-Syllable) strings. Error bars show standard error of the mean.

Error Bars: 95% CI

Error Bars: +/- 2 SE

The effect size for the difference between acceptance of Familiar-syllable XXY vs. New-syllable XXY was higher in the Single-Task group (Diff of Means = 0.38, *d* = 0.98) compared to the same difference in the Dual-Task group (Diff of Means = 0.28, *d* = 0.67).

Next, in order to probe the effect of *noisy-channel capacity* on rule induction, but crucially without an additionally overloading task, we compared the performance in the two conditions (Distractor and Single-Task) in a general linear mixed effects analysis of the relationship between Accuracy (correct acceptance of the grammatical test items and correct rejection of the ungrammatical ones) and Type of Test (Familiar-syllable XXY, New-Syllable XXY, Familiar-syllable X1X2Y, New-Syllable X1X2Y), Group (Distractor, Single-Task), as well as Group x Type of Test interaction. Therefore, as dependent variable we entered Accuracy score into the model. As fixed effects we entered Type of Test, Group and Type of Test x Group interaction. As random effects we had intercepts for subjects and items. 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. The model reported here is the best fitting model, both in terms of the model's accuracy in predicting the observed data, and in terms of AIC (Akaike Information Criterion).

We found a significant main effect of Type of test strings (F(3, 126) = 10.200, *p* < .001), and a significant Type x Group interaction (F(4, 83.311) = 2.517, *p* = .04). Although there was also a significant main effect of Group (F(1, 42) = 4.395, *p* = .04), we excluded it from the model because it did not improve it and it lead to an overfitted model.

Further analysis of the estimates of the main fixed effect of Type of test showed significant differences in the Accuracy scores (across both Distractor, Single-Task groups) for the Familiar-syllable X1X2Y (*t*(126) = -2.304, SE = .072, *p* = .02) and New-syllable XXY (*t*(126) = -3.528, SE = .072, *p* = .001) compared to the other types. The analysis of the estimates for the interaction effect Type x Group revealed a significant mean difference in the Accuracy scores between Groups (Distractor, Single-Task) for the Familiar-syllable X1X2Y (M = .159, SE = .065, F(1, 160.499) = 6.076, *p* = .01) and the New-syllable XXY (M = .141, SE = .065, F(1, 160.499) = 4.802, *p* = .03).

The effect size for the difference between acceptance of Familiar-syllable XXY vs. New-syllable XXY was higher in the Single-Task group (Diff of Means = 0.38, *d* = 0.98) compared to the same difference in the Distractor group (Diff of Means = 0.19, *d* = 0.55).

8. Discussion

The results of this experiment show that correct acceptance of the New-syllable XXY strings was higher when there was signal-irrelevant entropy (i.e. noise) added in the background, as compared to the Dual-Task condition, when participants were overloaded with an additional active task. Also, there was a difference between the rates of acceptance of New- vs. Familiar-syllable XXY depending on the experimental condition, which shows a difference between the groups in terms of how the XXY strings were encoded: the smaller the distinction learners make between a new and a familiar XXY, the more likely they are to have made the grammaticality judgement based on the *same-same-different* structure regardless of new/familiar syllable, i.e. *category-based generalization*. Thus, when comparing the correct acceptance of the New-syllable XXY strings to the correct acceptance of Familiar-syllable XXY strings, we found this difference to be smaller in the Distractor group as compared to the Dual-Task group (that is a smaller effect size of the difference in means between the groups).

Moreover, the correct rejection of Familiar-syllable X1X2Y strings was also higher when there was noise added in the background compared to when participants were overloaded with an additional task. The very low rate of correct rejection of Familiar-syllable X1X2Y in the Dual-Task group points to an impaired *item-bound generalization* due to the fact that participants' working memory was overloaded, thus it cannot flash out mismatches in combinations of specific items (i.e. familiar syllables) to help them reject these strings. Also, their *category-based generalization* is not strong enough to drive rejection of these strings based on a mismatch with a *category-based* XXY rule. Therefore, overloading working memory capacity with an additional task, which requires active switching of attention between two different tasks, results in the impossibility to keep in the focus of attention several familiar syllables and to bind them into regularities between specific familiar items, i.e. impaired *item-bound generalization*, or to bind them into categories, i.e. impaired *category-based generalization.*

Hence, taking all these findings into account, we conclude that when signal-irrelevant entropy (i.e. noise) was added in the background without an additional task, learners showed a higher tendency towards *category-based generalization*, than they did when they were overloaded with an additional task. These results support the hypothesis made by our model regarding the effect of a *noisy-channel capacity*: when disentangling the effect of taxing the working memory with an additional task from the effect of an attempt to exceed the channel capacity in purely information-theoretic terms (i.e. by introducing noise in the background), we found the tendency towards *category-based generalization* was higher.

When compared to the previous single-task experiment from Radulescu et al. (2019), in which there was no background noise, the correct acceptance of the New-syllable XXY strings was higher when there was signal-irrelevant entropy (i.e. noise) added in the background, as compared to the Single-Task condition (i.e. no added background noise, and no additional task overload). Also, when comparing the correct acceptance of the New-syllable XXY strings to the correct acceptance of Familiar-syllable XXY strings, we found this difference to be smaller in the Distractor group as compared to the Single-Task group (that is a smaller effect size of the difference in means between the groups). This difference shows that when signal-irrelevant entropy (i.e. noise) was added in the background without an additional task, learners showed a higher tendency towards *category-based generalization*, than they did when no noise was added in the background.

Moreover, the correct rejection of Familiar-syllable X1X2Y strings was also higher when there was noise added in the background compared to the Single-Task condition. These results support the hypothesis made by our model regarding the effect of a *noisy-channel capacity*: added signal-irrelevant entropy (noise) drives the tendency towards *category-based encoding*, regardless of the low target-intrinsic entropy in the input. However, in order to confirm this result, further research is needed with a larger sample and with control for participants' individual differences in memory capacity and visual pattern-recognition (RAVENS) in the single task also.

Although very weak, there was also a positive effect of Incidental Memory Task in the Dual-Task group, which showed that participants with a higher incidental memory capacity tended to have higher mean accuracy scores across Types of test strings, especially higher accuracy in the correct rejection of the Familiar-syllable X1X2Y. This finding shows that participants with a better incidental memory capacity were better able to incidentally remember the exact familiarization strings in order to confidently reject strings containing the familiar syllables but with a different structure.

9. General Discussion and Conclusions

The goal of this study was to probe the effect of the second main factor of our entropy model on rule induction, namely the effect of *noisy-channel capacity* on rule induction, by teasing apart the effect of overloading the "operational processor" with additional tasks from the effect of a *noisy channel* in information-theoretic terms. In order to do so, we employed the lowest entropy version (2.8 bits) of the XXY grammar we used in Radulescu et al. (2019), for which we found no evidence of *category-based generalization* (i.e. the acceptance rate for New-syllable XXY was at chance level, while the acceptance rate for Familiar-syllable XXY was at 95%). In this experiment, we presented adults with the same rule induction task on the same XXY language (i.e. the signal), thus keeping the low target-intrinsic *input entropy*, but we added stimulus-irrelevant entropy (i.e. noise) in the background. In one condition – the dual-task condition – we asked the participants to perform an additional memory task on the background noise material, while listening to the XXY language. In the other condition participants were exposed to the same XXY language with added background noise, but crucially they were not assigned an additional task on the background noise material.

The findings showed that learners' tendency towards *category-based generalization* is higher when exposed to a *noisy environment,* than it is when they are overloaded with an additional memory task simultaneously. Furthermore, when compared to our previous single-task "no noise" experiment from Radulescu et al. (2019), we found that a *noisy environment* drove *category-based generalization* despite the low language entropy, which did not support generalization in a noiseless environment. We interpret these findings to support the hypothesis of our entropy model regarding the effect of *noisy-channel capacity* on rule induction, namely that *noise* adds sufficient entropy, which increases the rate of equivocation and calls for a more efficient encoding method. This mechanism drives the need for a move to *category-based generalization*. Our findings are in line with the dynamic systems theory, according to which noise is a well-known catalyst for self-organizing into new structures (Prigogine & Stengers, 1984; Schneider & Sagan, 2005).

The first follow-up question for this model and findings about the *noisy-channel capacity* would be to define more precisely the *noise variable* in terms of the kind of noise that the model predicts to have a positive effect on the drive towards rule induction. More specifically, what kind of noise is predicted to have an effect on rule induction? Previous studies looked at the effect of *noise* from different points of view, for example, in terms of a specific kind of target-intrinsic entropy, e.g. noise determiners as opposed to main determiners in the determiner-noun pairs of an artificial grammar to be learned (Hudson Kam & Newport, 2009; Hudson Kam & Chang, 2009), or a noisy environment, that is added target-irrelevant noise, in the form of noisy location of stimuli or differently colored background in visual object-naming tasks (Stephen et al., 2009; Twomey et al., 2018), or a noisy source of information, i.e. less than perfect reliability of multiple intrinsic and extrinsic cues in a word-learning simulated task in a computational model (Monaghan, 2017). In our entropy model, based on the information-theoretic definition of noise (Shannon, 1948), we define the *noise* as any source of interference with the signal at either terminal ends of the communication system (i.e. either at the transmitter end, during the process of encoding the message into the signal, or at the receiver's end, during decoding the signal) or the interference in the channel, that is sources of noise that cause interference during transmission through the channel. Thus, at the computational level, any source that causes interference with the signal, such that it results in increased uncertainty (i.e. rate of equivocation) when decoding the signal at the receiver's end, is considered to be a source of noise.

More specifically, in the case of artificial grammar learning, sources of noise at the transmitter end could be envisaged as inconsistencies of the pseudo-artificial language system (i.e. the transmitter, according to the model we proposed in Radulescu et al., 2021), which encodes the message into the signal, i.e. the statistical properties of the signal that could be described as random (e.g. noise determiners, i.e. randomly occurring with any noun – Hudson Kam & Newport, 2009; Hudson Kam & Chang, 2009) or unreliable cues/features (e.g. less than perfect reliability of cues – Monaghan, 2017). Sources of noise during transmission through the channel are considered to be both internal, that is channel-intrinsic (e.g. properties and biases of the underlying cognitive capacities), and external, coming from the environment during transmission through the channel (e.g. stimulus-irrelevant background noise – Stephen et al., 2009; Twomey et al., 2018). Finally, sources of noise at the receiver's end which interfere with the process of decoding the signal to reconstruct the message, could be properties and biases of the underlying cognitive capacities involved in the decoding processes (e.g. working memory, interference from prior knowledge stored in the long-term memory), properties of the decision-making process, different learning strategies that learners employ in order to cope with the limited cognitive capacities, general biases regarding language composition and structure, etc.

At the algorithmic level, in order for the noise to create interference, according to the existent models of memory/attention (Baddeley, 2000, 2007, 2012; Baddeley et al., 2015; Cowan, 1988, 1995, 1999, 2005; 2016; Oberauer & Hein, 2012) and taking into account evidence from experiments with concurrent (interfering) tasks (Cocchini et al., 2002; Morey & Mall, 2012; Saults & Cowan, 2007), the noise material has to share physical properties with the target material to be learned, or the noise has to be somehow in the same domain as the target material.

The next question one might ask concerns the underlying cognitive capacities that would support an information encoding mechanism operating at the interaction between the *input entropy* and a finite *noisy-channel capacity*. Based on the description of processes and the findings of this paper, next we will briefly look into the assumed link between *channel capacity* and the underlying cognitive capacities hypothesized to be involved in rule induction. The finite encoding capacity (*channel capacity*) proposed by our model does not model in information-theoretic terms the limited cognitive resources, as modelled in classic resource-sharing models, thus a general resource-sharing ACT-R model (Anderson, 1993) will not be discussed or employed to account for the phenomenon under investigation. In ACT theory, the concept of capacity limitation is carried by the concept of activation levels, hence resource sharing in working memory is roughly defined as two concurrent tasks competing for the limited activation levels of elements in the declarative memory (Anderson, Reder & Lebiere, 1996).

The model proposed by Baddeley, with all its versions and additions starting from Baddeley and Hitch (1974) until Baddeley (2000, 2007, 2012) and Baddeley et al.(2015) focuses on the multi-component aspect of memory and looks at the interference between several components of memory. We deem compatible with our model some of Baddeley's concepts, namely, the central executive component which is defined as an attentional controller, rather than a memory component, responsible for attentional focus (i.e. directing attention to a specific task) and for dividing attention between several tasks. The central executive component could be envisaged to underlie what we dubbed the "operational processor" at the computational level. Moreover, the *episodic buffer*, added to the model to explain the link with long-term memory (Baddeley, 2000), with a proposed capacity of four chunks of information (Baddeley et al., 2015), seems also relevant as the component that allows linking and binding physical features of specific items, events into coherent episodes, which could be argued to underlie the *channel capacity*, in terms of the time-dependent finite amount of information that can be processed and encoded.

From a somewhat different perspective, Cowan's *embedded processes model* (Cowan, 1988, 1995, 1999, 2005) proposes a more attentionally-focused view on working memory (WM), by distinguishing between two components of

WM: (1) *activated* elements from *long-term memory* (LTM) and (2) the *focus of attention,* which contains a subset of the activated LTM. In this model, it is only the focus of attention which is actually limited to a number of separate chunks of information to be held in scope at one time, with Cowan (2005) arguing and showing that the working memory capacity is restricted to four chunks of information, rather than seven originally proposed by Miller (1956). So, in short, Cowan's approach to WM reflects an attentional capacity focused on a limited set of activated representations from LTM (these representations could be restricted by interference from other incoming items with similarities, and possibly by time-dependent decay). Our model seems to be very compatible with Cowan's model in that the limited *focus of attention* and the interference from similar items posed by Cowan's model could be envisaged to underlie, at the algorithmic level, the finite rate of information transmission (*channel capacity*) and noise interference of our entropy model.

A similar model, an extension of Cowan's embedded processes model, is the *concentric model* proposed by Oberauer (2002) and further developed in Oberauer and Hein (2012). The model proposes a concentric structure of three components with functionally separate regions: (1) the *activated part of LTM*, which might also serve for the retention of information in the short term, (2) the *direct access* which is hypothesized to hold four chunks of information available at a time, and to bind them into new structures, and (3) the *focus of attention* which singles out only one chunk to be used in the upcoming cognitive operation. The first two components are very similar to Cowan's model, and the limits of WM, as measured by several tasks (Cowan, 2001; Oberauer et al., 2000), are hypothesized to be restricted by the number of chunks of information that can be held in the *direct access* region, which corresponds directly to Cowan's *focus of attention*. Oberauer (2002) proposes that the capacity limitations are caused not by sharing limited resources, but by the challenge of selectively accessing several items that must be held available for cognitive operations in the *direct access* region. Therefore, Oberauer's model could also be compatible with our *channel capacity* model, with the *direct access* region being very similar functionally to our concept of *channel capacity*, and the *focus of attention* underlying what we dubbed the "operational processor", at the computational level. Moreover, in this model, increasing the amount of information in the *direct access* region leads to a slow-down of access to the particular items caused by interference between similar items, which relates to the hypothesis of our model that increased input entropy and/or noise drive the tendency to forget particular items and move to a more general *category-based generalization*.

By disentangling the effect of overloading the "operational processor" with an additional task from the effect of attempting to exceed the *channel capacity* with added entropy, our information-theoretic approach offers an explanation for previous apparently opposing findings. Hudson Kam & Newport (2009) found better generalization by overloading the memory with more target-intrinsic entropy in the learning material, which is in line with the prediction of our model that increased *input entropy* drives *category-based generalization*. Conversely, Perfors (2012) found no effect of working memory load on regularization during encoding by overloading the processing capacities with additional tasks at the time of learning, which is in line with the findings of this study. The *noisy-channel capacity* at the computational level maps onto what can be envisaged as an attentionally-taxed and error-prone encoding system with time-dependent limitations at the algorithmic level, and not to an overloaded task-handler.

Thus, at the algorithmic level, in accord with the *Less-is-More hypothesis,* we hypothesize that entropy (either input entropy or background noise) brings an inflow of information per unit of time in the working memory and distracts attention from the message (the signal). This distraction from the signal drives forgetting of the insignificant details, in order to prevent overfitting to existing past data for better generalization to future data. On the other side, additional tasks withdraw operational resources which are needed to be in place to bind the inflow of information into new structures. Thus, we hypothesized that the operational resources (which we dubbed the "operational processor") should not be overloaded with additional tasks. This was the rationale for designing two different conditions, in order to disentangle between the effect of the inflow of bits of information per unit of time, i.e. the source rate of information transmission, and the operational processor of WM that operates on the incoming bits of information. Our findings support this hypothesis, and show that switching between two active tasks places high demands on the *focus of attention* (Cowan's model) or on the *direct access region* (Oberauer's model), thus performance is overall worse in the dual task, than in the single task and the noise-added task.

To sum up, following suggestions from these models, we hypothesize that the cognitive capacities that underlie *channel capacity* at the algorithmic level are the attentionally-controlled regions of activated LTM (or working memory). It is important to mention that while all the current memory/attention models focus and account for several types of interference between capacities of memory components, and predict different levels of impairment of particular tasks (if there is interference from another task), none of these models predicts a better performance on a processing task as a result of another concurrent task or of additional incoming information into the focus of attention. In particular, our entropy model predicts better generalization, that is a transition from *item-bound generalization* towards *category-based generalization*, when an inflow of entropy (either target-intrinsic or background noise) attempts to exceed the *channel capacity*. This is a crucial part of our model which remains temporarily unaccounted for by the present models of memory/attention for encoding, in cognitive sciences.

This gap could be due to the particular focus of general memory and attention models on the faithfulness of memory representations (i.e. the persistence function of memory – Richards & Frankland, 2017), rather than on the property of the memories as models for future data/event integration and better generalization for the purpose of better adaptability to noisy environments (that is the transience function of the memory). Our model based on the *noisy-channel capacity* is very much in line with models from neurobiology, which propose an interaction between the persistence and transience functions of memory (Frankland, Köhler, & Josselyn, 2013; Hardt, Nader, & Wang, 2013; Migues et. al, 2016; Richards & Frankland, 2017), and also with neural networks research (Hawkins, 2004; Kumaran, Hassabis, & McClelland, 2016; MacKay, 2003). Specifically, these lines of research converge on the hypothesis that the memory system is designed with the goal of optimizing the method of encoding (or creating representations) such that future events/data can be more efficiently integrated in the representations, i.e. for better generalization and prediction of future data/events, in order to allow for more flexibility and better adaptability to noisy environments. More precisely, the above-mentioned converging views and evidence from neurobiology and neural networks show that our memory system encodes representations in such a way to prevent both underfitting (i.e. to prevent forgetting relevant parameters which help correctly capture the underlying data structure), and overfitting to past data/events (that is to prevent incorrectly remembering and encoding noise as underlying structure). Both in neurobiology and in neural networks research, noise injection (i.e. adding random variability to synaptic connections – Hinton & van Camp, 1993) is used, among other techniques, as a means to prevent overfitting to past data, which in turn promotes better generalization to novel input in noisy environments (Richards & Frankland, 2017). In accord with these current developments in neurobiology and neural networks research, our model proposes the *noisy-channel capacity* to reflect and quantify, at the computational level, this design feature of the memory system proposed in neurobiology and neural network research that naturally and automatically acts as a sweet spot between under- and overfitting to past data, i.e. creating memory representations as efficiently predictive models of novel data. As per our model, we argue that the mechanism at stake, from an information-theoretic point of view, is that noise adds enough randomness (=bits of entropy) in the data which is higher than the upper bound of the degree of details allowed by the *channel capacity,* and results into a high rate of equivocation, which in turn creates the need for another more efficient encoding method, in order to avoid exceeding the *channel capacity*. This is the main contribution that our information-theoretic model adds to this line of research.

Another relevant question that one might ask would be why should added entropy (be it input entropy or stimulus-irrelevant noise) *drive a need* to find a more efficient encoding method, at the computational level? Shannon's channel capacity theory posits that an increase of the rate of equivocation renders the encoding method inefficient, and that "it is possible to find another encoding method", but it is not possible to exceed the actual rate of transmission of information, i.e. *channel capacity.* Since information theory is about electrical communication systems, not about biological systems *per se*, one might raise the point that it does not offer a direct explanation as to what *drives the need* to find another encoding method for our biological encoding system. Therefore, in order to answer this question, we extend the entropy model further by linking it with the dynamic systems hypothesis which is relevant to self-organizing systems, where entropy (and noise) is a driving force towards new structures. This link was suggested in other studies in cognitive science (Stephen et al., 2009) and it makes sense theoretically as well, since although Shannon's entropy formula was devised in order to quantify the amount of information when transmitting an electrical system, it is basically the same as Boltzmann's formula for thermodynamic entropy (Karnani, Pääkkönen & Annila, 2009; Plenio & Vitelli, 2001; Trambouze, 2006), which applies to all biological systems (Prigogine & Stengers, 1984). In dynamic systems theory, self-organization, as a natural property of complex systems, was offered as an account for the emergence of new structures (Prigogine & Stengers, 1984; Schneider & Sagan, 2005): the configuration of the constituents of a system will remain unchanged until an increase in entropy overwhelms its internal boundaries and the system approaches a critical instability where the constraints between internal parts dissolve, thus setting them free to interact and bind into a new configuration spontaneously. Hence, in self-organization, new structures are predicted by an increase in entropy. However spontaneous self-organization is, this should not be confused with a random re-structuring of the constituent parts. Self-organization occurs with the precise purpose of rendering the system into a better dissipative structure for entropy, that is a structure which is better fitted to dissipate more entropy more efficiently. Here lies the link with our entropy model, in terms of the need for a more efficient encoding method, that is a new structure that will allow the rate of transmission of information to reach its maximum, that is to reach the maximum amount of entropy that can be transmitted per unit of time with the least rate of equivocation (i.e. highly efficient).

Our entropy model offers an extended and fine-grained information-theoretic approach to the *Less-is-More hypothesis* (Newport, 1990) at the computational level, and agrees with findings from neurobiology (Frankland et al., 2013; Hardt, Nader, & Wang, 2013; Migues et al., 2016; Richards & Frankland, 2017), and from neural networks research (Hawkins, 2004; Kumaran et al., 2016; MacKay, 2003), which converge on the hypothesis that the memory system (and thus the neural network modelling) is designed for optimized generalization and decision-making, by having the capacity to encode a finite degree of specificity (i.e. entropy, in information-theoretic terms), in order to prevent overfitting to past data and to allow for future adaptability to noisy environments. Our findings and model also agree with the theory of dynamic systems (Prigogine & Stengers, 1984; Schneider & Sagan, 2005), which suggests that self-organizing, complex systems remain unchanged until entropy (or noise) overwhelms the order of their internal constituents, such that old bindings dissolve and free up the constituents causing them to interact and re-bind into new structures. Our model and findings tap into the interaction between the inflow of entropy (and noise) and the design features of our encoding system, and thus, they add to the research on the dynamical processes that drive self-organization for higher-order cognitive phenomena (Stephen et al., 2009).

Appendix A

Familiarization strings
keːkeːmy
jujuɣo
daːdaːli
pypyveː
tøːtøːrøː
hihisaː
fofoʃu
keːkeːmy
jujuɣo
daːdaːli
pypyveː
tøːtøːrøː
hihisaː
fofoʃu
keːkeːmy
jujuɣo
daːdaːli
pypyveː
tøːtøːrøː
hihisaː
fofoʃu
keːkeːmy
jujuɣo
daːdaːli
pypyveː
tøːtøːrøː
hihisaː
fofoʃu

Appendix B - Test strings

Test 1		Test 2	Test 3	Final Test	
Familiar-syllable XXY	daːdaːli	hihisaː	keːkeːmy	tøːtøːrøː	jujuɣo
New-syllable X1X2Y	poxaːru	runyni	xaːmisy	syniny	mininy
New-syllable XXY	dydytaː	zuzuvo	sosory	jijiføː	ʋuʋuseː
Familiar-syllable X1X2Y	judaːsaː	pytøːmy	keːfoveː	hidaːrøː	tøːpyɣo

Chapter 7

Directions for Future Research: Towards a Thermodynamic Theory of Rule Induction

Radulescu, S. and Avrutin, S.[31]

The aim of this chapter is to suggest directions for future research on rule induction by laying the foundations of a new theoretical framework based on an innovative thermodynamic model of rule induction. We think that a comprehensive theory of rule induction should be built on biologically plausible mechanisms, and formulated in accord with the laws of biophysics and neuroscience. Information theory provides a straightforward bridge between these fields of study, by employing entropy-related concepts that are ultimately linked to the same concepts in biophysics. In this dissertation, we proposed an information-theoretic entropy model, and showed in several artificial grammar studies that rule induction in language is an information encoding mechanism resulting from the brain's sensitivity to increasing information *entropy* interacting with the *channel capacity*, that is the brain's finite rate of encoding information. But *why* is the brain sensitive to information entropy? Further, *why* and *how* does rule induction emerge? Information entropy is a reflection of thermodynamic entropy, and they are to a large extent equivalent (Karnani, Pääkkönen & Annila, 2009; Le Bellac, Mortessagne, & Batrouni, 2004; Sethna, 2006). Recent studies in biophysics, and biosciences in general, converge on a thermodynamics view on the brain as an open dissipative system that operates under the rule of the laws of physics, focusing on cognition and consciousness as being affected and driven by the laws of thermodynamics (Annila, 2016a, 2016b; Collell & Fauquet, 2015; DeCastro, 2013; Del Castillo & Vera-Cruz, 2011; La Cerra, 2003; Sharma & Annila, 2007; Varpula, Annila, & Beck, 2013; Yufik, 2013). Here we propose a framework (and briefly sketch a model) that connects a thermodynamic cognitive scientific view on rule induction with the information-theoretic model proposed in this dissertation and other information-theoretic cognitive models (Friston, 2010).

Firstly, this section reviews the previously proposed thermodynamic models of cognition together with information-theoretic perspectives on brain activity, in order to explain the link between information-theoretic concepts, like information, information entropy, and thermodynamics concepts, like thermodynamic entropy and free energy, as well as their relevance for cognitive

[31] This chapter is a modified version of a manuscript in preparation: Radulescu, S., & Avrutin, A. (2021). Towards a Thermodynamic Theory of Rule Induction

processes. This link is crucial to understanding a thermodynamic theory of cognition, in general, which was already proposed by some early studies (Kirkaldy, 1965), and coined as "cognitive thermodynamics" (Yufik, 2013) or as a "general thermodynamic theory of cognition" (Annila, 2016a; Varpula et al., 2013), or as a "simple general principle of brain organization" (Velazquez, Mateos, & Guevara Erra, 2019). The review lays the foundations for proposing a specific thermodynamic and information-theoretic model of the cognitive system with a particular focus on the process of linguistic (as well as, general) rule induction.

The review of the above-mentioned studies will be organized based on the following four questions/problems that we have identified as relevant to a thermodynamic theory of cognition in general, and of rule induction, in particular. A legitimate first question about a possible link between thermodynamics and cognitive (information-processing) mechanisms would be related to what appears to be a counter-intuitive idea in psycholinguistics and cognitive sciences in general: information, which is regarded as something "abstract", as well as cognitive processes, would have to be "concrete" or physical in nature in order to be shown as governed by the laws of thermodynamics. Thus, firstly we will show the physical nature of information, which is a widely accepted and already basic concept in physics and biophysics (Annila, 2016a; Brillouin, 1953; Karnani et al., 2009; Landauer, 1961; 1991; Plenio & Vitelli, 2001).

Secondly, another myth had to be debunked in biosciences in order to understand how the laws of thermodynamics, especially the 2nd law, should be applied to open systems, like the brain (i.e. which interacts with the environment). Bejan (2017) explains that the misconception was related to the fact that the 2nd law was traditionally explained and taught by physicists in relation to heat engines and as a law that applies to isolated systems (i.e. which do not interact with the surroundings). Recent increasing attention has been given to the application of the 2nd law of thermodynamics to non-isolated (open) systems, among which biological systems (Annila & Beverstock, 2016; Avery, 2012; England, 2013; 2015; Sharma & Annila, 2007).

Thirdly, regarding a specific entropy-related formulation of the 2nd law of thermodynamics, there has been considerable misunderstanding within cognitive sciences, biology and even physics, which was caused by a description of entropy as "disorder of the system". Hence, a particular formulation of the 2nd law as a natural tendency of all things to flow from order (i.e. less entropy) to disorder (more entropy) has created a misleading view. "The entire universe is collapsing into disorder," has been the gloomy prediction ever since. This overly simplified formulation contravenes the general obvious tendency of life and things towards patterns and structure, which is the opposite of disorder. As a result, there was an apparent dissonance between views in biology and biophysics (Schrödinger, 1944), which argue that life tends towards structured forms (i.e. less entropy), as opposed to the standard views of physics and geophysics, which hold the universal tendency towards higher entropy, as a more probable state, as per the 2nd law. As a consequence, since open dissipative systems were described as spontaneously creating entropy-dissipative structures (i.e. reduced-entropy structures), seemingly in violation of the 2nd law, researchers advanced proposals regarding open systems (Nicolis & Prigogine, 1989; Prigogine, 1978). Specifically, it was proposed that system-internal entropy should be reduced at the expense of increasing the entropy dissipated across the boundaries of the open system. While this was a legitimate problem, the tendency of open systems to evolve towards structure has been recently shown not to be an actual violation of the 2nd law (Annila & Baverstock, 2016; Avery, 2012; Bejan, 2017; Bejan & Marden, 2009; Bejan & Zane, 2012; Sharma & Annila, 2007).

And finally, a recent ongoing debate among physicists, which is relevant for this discussion, concerns the question whether the tendency towards structure, and the evolution of life in general, can be fully explained and derived from the 2nd law of thermodynamics. In particular, recent developments in thermodynamics, that is in the late 20th century, proposed and showed evidence for yet another law of thermodynamics, namely the constructal law. This law is argued to account for the evolution of structure (i.e. configuration, design) of everything in nature, from inanimate to animate systems, and from natural to man-made structures (Bejan, 1996, 1997a – d, 2012; Bejan & Lorente, 2004; 2010).

Thus, in accord with the latest developments in thermodynamics, we propose that thermodynamic models of the brain and cognition, in general, as well as our specific thermodynamic model of rule induction, should take into account both the 2nd law of thermodynamics and the constructal law. The constructal law predicts and explains several other descriptive principles and empirical laws that were previously posed both in biology and cognitive sciences (e.g. the principle of entropy reduction and the maximum entropy hypothesis – Bejan & Zane, 2012) and in linguistics and psycholinguistics, e.g. the principle of least effort (Zipf, 1949) – Zipf's law (Bejan & Zane, 2012). Further, our model based on the thermodynamics laws could predict and explain several entropy-related empirical principles that were proposed in psycholinguistics: entropy-reduction based models of language comprehension and sentence processing (Hale, 2006; Levy, 2008; Linzen & Jaeger, 2016; Venhuizen, Crocker, & Brouwer, 2019a; 2019b) and entropy-reduction based accounts of rule learning and regularization (Ferdinand, Kirby, & Smith, 2018). So far, to the best of our knowledge, entropy-based accounts in psycholinguistics have only adopted an information-theoretic perspective on entropy, but not a thermodynamics perspective.

The thermodynamic model of rule induction proposed (briefly) in this chapter aims at explaining the mechanisms of information encoding underlying rule induction as natural automatic processes sustained by the brain's neural networks. The neural networks underlying information encoding are designed by the natural laws of thermodynamics for the ultimate purpose of facilitating the flow of energy or consumption of free energy[32] in the most efficient way, i.e. in the least possible time (Varpula et al., 2013). According to the thermodynamic model proposed here, rule induction is a natural reaction of the neural networks to consume free energy, and consequently dissipate[33] entropy in the environment, by creating structure under the governance of the laws of thermodynamics.

Rule induction happens to us when exposed to language (or an influx of information, in general), just like photosynthesis happens to a flower in the sunlight. And this is not a metaphor, but a physical and real process. It should be clarified from the beginning that our intention is not to propose a specific mathematical thermodynamic model of neural networks, but rather a theoretical framework to inspire an innovative account on rule induction, as an information processing and encoding mechanism under the governance of the laws of physics. We suggest that this research direction should be named thermodynamic psycholinguistics.

We propose the first joint information-theoretic and thermodynamics perspective on rule induction. This new perspective suggests that the 2nd law of thermodynamics can answer the question *why* rule induction happens, while the constructal law of thermodynamics can answer the question *how* rule induction happens. Specifically, according to the 2nd law of thermodynamics, we propose that rule induction happens as a natural result of the tendency of our brain's neural networks (and consequently, our cognitive system) to consume free energy (in the form of information) in the least time possible. The constructal law predicts the generation of particular evolving structures (design, configurations) that facilitate the flow of energy (which drives efficient consumption of free energy or efficient information transmission).

Using the constructal law, we suggest that rule induction is a flow system, as part of the bigger flow system which is language. As defined in the constructal law, everything that moves – animate or inanimate – is a flow system. Flow is defined as the movement of an entity relative to another, i.e. a current or a stream originating from a point and moving to other points (Bejan & Zane, 2012). We hypothesize that, just like all the other flow systems in nature, rule induction has evolved for the purpose of facilitating faster and better flow (or transmission) of information. Thus, we suggest that the constructal law predicts the tree-like hierarchical structure of language, just as it predicts the tree-like hierarchical structure of other flow systems (Bejan & Zane, 2012). Further, the constructal law predicts the design of rule induction (and categorization) according to the constructal principle of few large channels of energy dispersal and many small channels (Bejan & Zane, 2012): a few large channels for information flow – few general categories (via *category-based generalization)* and many small channels – many specific items and item-bound relations (*item-*

[32] Free energy is the amount of energy that can be used to produce useful work, as opposed to entropy (Schrödinger, 1944).

[33] Dissipate means to lose (energy, such as heat) irreversibly.

bound generalization), because this is the most efficient way for information to flow, i.e. to be transmitted, in information-theoretic terms. This particular design feature of rule induction cannot be predicted by an information-theoretic model alone. Thus, we propose that further research into rule induction should follow this joint information-theoretic and thermodynamic framework/model in order to test these hypotheses and predictions. This framework/model renders unnecessary the ad-hoc postulation of psycholinguistic mechanisms and cognitive scientific principles, which lack a plausible biophysical foundation and/or neurobiological evidence.

1.1 Information is physical

What is information? Although many scientists agree that information is a fundamental property of nature, and cognitive sciences describe the brain as an information processing machinery, definitions of information are vague or intuitive, at large, and they differ depending on the field of application (Pepperell, 2018). Colloquially and by the dictionary definitions, information means knowledge, data or facts learned, or intelligence (news) acquired about something.

Mathematically, the first precise definition of information was given by Shannon (1948) in his theory of communication:

$\mathrm{I}(x_i) = -\log_2 p(x_i),$

where (x_i) is a symbol in a transmitted message and $p(x_i)$ is the probability of occurrence of that symbol. Base 2 of the logarithm defines the bit as the unit of measurement for information. It is obvious then that information is defined in information-theoretic terms as a function of probability of occurrence of something, and does not have to do with the meaning attached to symbols. Shannon (1948) defined the average information content of a message containing n symbols, each with its probability of occurrence $p(x_i)$ as entropy (H):

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

Intuitively, Shannon's entropy, which became known as information entropy, quantifies the average uncertainty per symbol carried by a message, or the average variability of the message, as it varies both depending on the number of symbols and their probability distribution.

By these definitions, concepts of information content and average uncertainty per symbol in information-theoretic terms sound quite abstract. However, the physical nature of information might not necessary come as a surprise in the age of computers and digitalization, when it is common knowledge that information is stored, processed and transmitted by physical means, such as hard drives, processor chips and optical fibers. All these devices, their functionality and efficiency, are governed, empowered or limited by the laws of physics (Plenio & Vitelli, 2001).

In order to specify the link between information and informational entropy and concepts that are commonly regarded as "physical", not "abstract", like energy and thermodynamic entropy, which will eventually shed light on how the mind and cognition are intrinsically energy consumption processes, we will briefly introduce a long-standing and thought-provoking physics puzzle – Maxwell's demon. The solution to this problem brought the first intuitive insight into the close relation between information, energy used for work and the laws of thermodynamics.

Maxwell's demon is a thought experiment proposed in 1867 by the physicist James Clerk Maxwell, which was conceived to pose a question in relation to the 2nd law of thermodynamics. In short, the 2nd law of thermodynamics states that, left alone, heat always flows from hot objects to cold objects, pressure falls from high to low. The most obvious example is the hot cup of tea left on the kitchen table which becomes cold in a couple of hours, while a glass of cold coffee does not become iced coffee after being left on the same kitchen table. Both liquids will reach what we know to be the room temperature. In time, nature tends towards uniformity. If we want to reverse the process we have to put some effort, that is to do some work and spend some energy to either warm up the cup of tea or to ice-cool the coffee. This is a law of nature: we have to pay the energy bill, if we want to go against the natural flow of heat from hot to cold.

Now, going back to Maxwell's thought experiment, imagine a situation where we would want to build an oven next to a fridge in the kitchen to keep the cup of tea hot next to the iced coffee, without paying the energy bill. In other words, imagine a transparent container divided by a wall in two chambers, A and B, both having the same air temperature in the beginning of the experiment. Temperature in physics is given by the kinetic energy or velocity of the air molecules (i.e. the speed and direction of the molecules movement).

A demon, who stands guard on top of the container, monitors the system, namely the speed and direction of each molecule. He must probably be equipped with a laser or electron device to bounce light or electrons off the moving molecules in order to measure the speed and observe the positions of the molecules. His goal is to use the information about the speed and direction of the air molecules, in order to obtain a fridge in chamber A next to an oven in chamber B, without paying the energy bill, i.e. without using any energy, only information. So, he uses the information for his purpose as follows: he opens a trapdoor in the wall when he sees a fast-moving (=hot) molecule heading towards the right chamber (B) to let the fast molecule travel to B. But when he sees a slow-moving (=cold) molecule traveling towards the left chamber (A), he opens the trapdoor to let the slow molecule go to A. In time, the demon creates a cold chamber (A) next to a hot chamber (B): A has an average lower molecular speed than B. And this was obtained only by observing the molecules, i.e. acquiring information about the molecules (fast- or slow-moving): apparently without paying the energy bill, the demon laughs in the face of the 2nd law of thermodynamics.

The demon managed to tamper with the natural flow of heat (i.e. energy) towards uniformity, without spending any energy, only by using information about the system. Before presenting the final solution to this problem, which actually came after a century of heated debates, and without having any insight into the concepts of thermodynamics, the reader can intuitively understand from this clever thought experiment, that information can somehow be transformed into useful work. This was the beginning of understanding the physical nature of information.

Szilard (1929) provided a first solution to the problem, and hence brought the first insight and proof of the intrinsic relation between information and energy, by showing that obtaining information from a system always involves an energetic cost of *kTln2 J*[34] per bit of information, and thus an increase in the heat (energy) irreversibly released in the world of *kln2 J/K* per bit of information. This irreversible dissipation of energy in the environment is the increase of thermodynamic entropy in the world, which is a consequence also covered by the 2nd law, and it will be further addressed in the next section. Further, Landauer (1961) demonstrated that when information is erased, there is always a minimal energy cost of *kTln2 J* per bit of information, hence always a minimal amount of heat (energy) irreversibly dumped in the environment, which equals *kTln2* per bit of information.

Going back to Maxwell's puzzle, intuitively, the demon has a limited memory system where he stores information about the speed and direction of movement of the millions of molecules in order to be able to track them and perform the informed decision when to open the trapdoor. The information derived from observing the positions and movements of the molecules has to be stored in the demon's memory system, and the information (measured by Shannon's entropy) in the mind of the demon is a reflection of the entropy of the container's state, given by all the possible positions and velocities of the molecules (as per Boltzmann's definition[35]). As such the entropy in the mind of the demon has increased to a higher extent than the entropy of the container has decreased by forcing each molecule into one of the container sides. Also, in the process, old recordings of information have to be erased from his limited memory in order to make room for new recordings of information about the molecules. In order for the demon's memory system to perform all the measurements and erase old information in a timely and ordered manner, he needs quite some computation power, which cannot come for free without paying the energy bill. Per Landauer's principle (Landauer, 1961), for every bit of erased information there is an energy cost associated with dissipation of energy in the environment and increasing entropy in the environment. According to calculations, at the end of the process, the whole system of the

[34] *k* is the Boltzmann constant, approximately 1.38×10^{-23} J/K (Boltzmann; 1877).

[35] Boltzmann (1877) defined thermodynamic entropy as a state of the physical system given by the number of possible positions and velocities of gas molecules: in short, low possible positions and velocities describe low entropy states, while high possible positions and velocities describe high entropic states.

demon's computation and memory system together with the container will have dissipated more entropy in the environment than the entropy that was reduced inside the container by the demon's sorting.

Later, Bennet (1982) looked at the same problem in computation and showed that measuring information can in principle be performed reversibly, that is without an increase in the energy dissipated in the environment, but erasing the previous measurements to make room for the next measurements cannot avoid the energetic cost, thus the generation of entropy. Bennet (1982) made the argument that the processed information has to be stored and encoded in a physical medium which must obey the laws of physics, in this case the demon's memory, and thus there is also a minimal energetic cost that must be paid for encoding one bit of information, which is also equal to *kTln2 J*.

The solution to Maxwell's demon shows that information is physical, not "abstract" or a purely mathematical concept, and that in order to process and encode information, in this case specifically observing (fast vs slow molecules), encoding the information (on speed and position of molecules) and erasing of information (from the memory of the demon to make space for the next measurement), energy must be spent, and as a consequence the processing and encoding of information causes generation of entropy in the environment. Roughly speaking, the demon spends *kTln2 J* to measure if one molecule is fast or slow, and encodes this as one bit of information, which subsequently has to be erased by spending another *kTln2 J.* Finally, Maxwell's demon puzzle was solved, at least theoretically, and the 2nd law of thermodynamics was reinstated as ubiquitous. Since recent technology allows for experiments with atoms and particles, Maxwell's demon and its proposed solution were probed and demonstrated in laboratory experiments (Price, Bannerman, Viering, Narevicius, & Raizen, 2008; Raizen, 2009), as well as Landauer's principle (Bérut et al., 2012).

In conclusion, the apparent violation of the 2nd law of thermodynamics by Maxwell's demon was solved by the remarkable insight into the intrinsic relation between information and energy. This insight shows that information is physical and thence, any way of processing and encoding information comes with an energy cost and, consequently, an increase of entropy in the world. Having shown that information processing and encoding has a minimal energetic cost in natural systems and in computation alike creates the necessary premises to state that information processing and encoding by the brain during cognitive processes, like learning, reasoning, etc., must be accompanied by an energetic cost and, thus an increase of the thermodynamic entropy in the world (Karnani et al., 2009). The exact energetic costs involved in cognitive processes, and in fact any brain processes, have yet to be thoroughly investigated, although there are some studies in this direction in some areas, e.g. in cellular computation (Mehta & Schwab, 2012), in vision of blowflies (Laughlin, van Steveninck, & Anderson, 1998).

Summarizing, based on these premises and the illustrative case of Maxwell's demon, any way of processing, erasing and encoding information involves energy consumption, and as a result an increase of entropy. Hence, since rule induction – which draws on memory resources – involves processing, erasing and encoding information, we conclude that (linguistic or general) rule induction is an energy-consuming mechanism governed by the 2nd law of thermodynamics, which generates an increase of thermodynamic entropy in the world.

1.2 Second law of thermodynamics for open systems

The earliest formulation of the second law of thermodynamics was given by Sadi Carnot in the early 19th century, but this formulation did not include the concept of entropy explicitly. It was used in the formulation of the maximum efficiency of heat engines, by describing how heat can be transformed into work. In 1856, Clausius noticed a universal principle built into everything in the universe: heat always flows into one direction, from hot to cold objects. As heat (or energy) always flows from hot to cold, entropy (S) always increases:

$$\frac{dS}{dt} \geq 0.$$

This is the first entropy-related formulation of the 2nd law of thermodynamics[36], and it was formulated in relation to heat engines and stated about isolated systems, i.e. systems that do not exchange heat (energy) or mass with the environment. This is the reason why a short easy-to-memorize formulation of the 2nd law has commonly been used since then: the entropy of an isolated system will always increase (Collell and Fauquet, 2015; Prigogine, 1978).

However, although correct about the isolated systems, this is not the most accurate formulation of the law (Bejan, 2017; Feynman, Leighton, Sands, & Hafner, 1965). The 2nd law of thermodynamics states that if work is done against friction, the work done is equal to the heat produced, and that the heat produced cannot be changed back into work (hence, the irreversibility of the process) *with no other change in the system or its surroundings*. In other words, in *irreversible* changes, the entropy of the system and of the whole world always increases. Only in reversible processes, does the entropy stay constant, and since no process is truly reversible, there is always at least a small increase in entropy in the world (Feynman et al., 1965).

The application of the 2nd law of thermodynamics to non-isolated (open) systems, among which biological systems, has received increasing attention recently (Annila & Beverstock, 2016; Avery, 2012; England, 2013; 2015; Sharma & Annila, 2007). Rephrasing the 2nd law of thermodynamics in modern terms:

[36] Since the reader might want to be reminded of the 1st law of thermodynamics, although not immediately relevant to the topic of this chapter, here it is: the 1st law of thermodynamics is the principle of conservation of energy, and it states that the energy of the universe is always constant. In its most accurate formulation: if one has a system and puts heat into it, and does work on it, then its energy is increased by the heat put in and the work done (Feynman et al., 1965).

spontaneously, energy always goes from being concentrated to being spread out or dispersed.

Recent research in the physics of animate and inanimate life forms converge on the idea that, as per the thermodynamics of the open systems, the $\frac{dS}{dt}$ rate acts as a natural selection criterion that chooses mechanisms and structures (organisms) that are better and faster at taking in energy from the surroundings and facilitating the energetic flow in order to increase entropy, i.e. to diminish the amount of energy available for doing work – *free energy* (Annila & Annila, 2008; Avery, 2012; Bejan, 1997; Bejan & Zane, 2012; England, 2013, 2015). Rephrasing the 2nd law of thermodynamics in terms of the tendency to consume free energy, or in other words, level off the gradients of energy between the organism and its surroundings by energy transduction (transformation of one form of energy into another form of energy) could inform the thermodynamic models of the brain, and consequently of the information processing mechanisms.

Thus, adding to the conclusion of the previous subsection, rule induction (linguistic or general) can be envisaged as an energy-consuming mechanism governed by the 2nd law of thermodynamics as it applies to open systems: as a reflection of nature's evolution towards a dissipation-driven mechanism that facilitates leveling off the gradients of energy (in this case, driven by information processing and encoding).

1.3. Thermodynamic entropy and information entropy

The laws of thermodynamics govern the amount of *available* energy and they involve the concept of entropy for irreversible thermodynamic processes. Although routinely defined as the quantification of the degree of disorder of a system, *entropy,* as per its precise thermodynamic definition, is the amount of energy dissipated as molecular vibration that cannot be used to produce work (Feynman et al., 1965).

The first definition of thermodynamic entropy was formulated by Clausius in 1856 as:

$\Delta S = \frac{\Delta Q}{\mathrm{T}}$,

where ΔS stands for the change in entropy, ΔQ stands for the change in heat and *T* for the absolute temperature of the system, and thus physical entropy is measured in Joules/Kelvin (Feynman et al., 1965).

While Clausius gave a macroscopic definition of entropy in terms of heat change, Boltzmann (1877) gave a microscopic definition of physical entropy, from the perspective of statistical mechanics:

S = *k*ln(W), where *k* is the Boltzmann constant (approx. 1.38 x 10^{-23} J/K), and W represents the number of equiprobable microstates of a system. In detailed form, if counting the number and different probabilities of microstates, Boltzmann's formula for entropy (S) differs only by a constant (*k*ln2) from Shannon's entropy formula:

$\mathrm{S} = -k\ln 2 \sum_{i=1}^{n} p_i(\log p_i)$.

Also Gibbs formula for entropy has a similar form to Shannon's entropy, with the same difference of a change in units given by Boltzmann's constant *k*:

$$\mathrm{S} = -k \sum_{i=1}^{n} p_i (\ln p_i)$$

Based on this formula equality, it was shown that theoretically in order to convert a bit of information entropy in units of thermodynamic entropy one has to multiply the information entropy by the constant *k*ln2 (Plenio & Vitelli, 2001). As per Landauer's principle (Landauer, 1961), this is theoretically the thermodynamical entropy generated by erasing one bit of information. Also, it was shown that any operation of processing information, e.g. erasing or encoding one bit of information, making a yes/no decision, and in principle any logically irreversible computation (Lutz & Ciliberto, 2015), results in the same thermodynamical entropy generated per bit of information in the world (Bennet, 1982).

Based on these considerations, among others, and on shared mathematical properties (Karnani et al., 2009), information entropy can be regarded as a reflection of thermodynamic entropy, given that they are to a large extent equivalent (Karnani et al., 2009; Le Bellac et al., 2004; Sethna, 2006).

Entropy has often been regarded as the disorder of a system, however nature's tendency towards highly organized structures seems to point towards a resistance to increasing disorder. According to recent views (Annila & Beverstock, 2016; Bejan & Zane, 2012), the misconception of entropy as being equal to the disorder of a system, and of the 2nd law of thermodynamics as being a gloomy prediction of the collapse of the world into disorder stemmed from simplistic erroneous formulations of Boltzmann's entropy, and it has created a lot of unnecessary confusion (Sagan, 2008).

Boltzmann (1877) described a low number of possible positions and velocities of gas molecules, i.e. low entropy, as reflecting an orderly arrangement of gas molecules, while increasing the number of possible positions and velocities of molecules, i.e. high entropy, as describing a disorderly state. As a result, conceptualizations of entropy as disorder of a system, and increasing entropy as increasing disorder led to a need to postulate solutions to account for life's obvious tendency towards order and organization (Avery, 2012; Annila & Beverstock, 2016; Prigogine, 1978).

Among these solutions, Prigogine and colleagues' proposal (Nicolis & Prigogine, 1989; Prigogine, 1978) for nature's tendency to evolve towards ordered dissipative structures has been of great influence both in biophysics and in cognitive sciences: they propose that the entire entropy of a system should be seen as two terms, an internal entropy (dS_i/dt) generated inside the system and an external entropy (dS_e/dt) exchanged through the boundaries of the open system. When there is an increase of internal entropy (dS_i/dt) beyond the limit that the current organization or structure of the system can dissipate, this gain in entropy must be immediately released through the boundaries of the open system into the environment. This need was proposed to drive self-organization of natural open systems into new structures, which are better dissipative structures (Prigogine, 1978). This approach was adopted in thermodynamic models of brain activity, which regard the brain as an evolving structure towards better dissipation of entropy through its boundaries (Del Castillo & Vera-Cruz, 2011; La Cerra, 2003), and in cognitive processes such as cognitive reorganization during a problem-solving task by an initial increase followed by a drop in internal entropy (Stephen et al., 2009).

To summarize, thermodynamic models of brain activity and cognitive processes have mainly proposed that the brain tends towards reducing entropy, by creating new structures better adapted for entropy dissipation. In line with this approach, although not directly inspired by it, but based on information entropy, similar views from psycholinguistics and cognitive sciences proposed entropy-reduction based models of language comprehension and sentence processing (Hale, 2006; Levy, 2008; Linzen & Jaeger, 2016; Venhuizen et al., 2019a; 2019b) and entropy-reduction based accounts of rule learning and regularization (Ferdinand et al., 2018).

However, another view on the 2nd law of thermodynamics and especially on entropy from its more accurate definition from the perspective of energy dispersal (Feynman et al., 1965) derives entropy from the physics of open systems and formulates it in terms of consumption of free energy: the principle of increasing entropy of the 2nd law equals the imperative to decrease free energy (Annila, 2016a; 2016b; Annila & Beverstock, 2016; England, 2013, 2015; Sharma & Annila, 2007; Varpula et al., 2013). Thus, living organisms and the brain are proposed as energy-consuming structures governed by the 2nd law of thermodynamics, such that they evolve towards consuming free energy in the least possible time, which in turn drives increasing entropy dissipation. Mounting proposals have been advanced in this direction of the free energy principle as the underlying first principle of living organisms (Colombo & Wright, 2018; Friston 2009; Friston & Stephan, 2007).

Therefore, another concept that becomes relevant to this discussion is the concept of *free energy* (or *negentropy*). The concept was introduced by Schrödinger (1944) in relation to a thermodynamic perspective on living organisms, and dubbed negative entropy, because it is the energy that is available to produce work, as opposed to entropy. Free energy can be described either by Helmholtz free energy formula: $F = U – TS$, or by the standard Gibbs free energy equation $G = U - TS$, where U is the internal energy of the system, T the temperature and S the entropy (Feynman et al., 1965). According to Gibbs free energy formula, when a chemical process takes place, heat is exchanged with the environment, and a process is spontaneous if it entails a decrease of the Gibbs free energy.

Moreover, it was shown that Gibbs free energy is a measure of the "thermodynamic information" contained by a system (Avery, 2012). In fact, information and negentropy (or free energy) were first shown to be interchangeable thermodynamic quantities by Brillouin (1953). Brillouin showed that information corresponds to a negative term in the final entropy of a physical system:

$S_1 = S_0 - I,$

where S_0 is the initial entropy of a physical system, I is the information about the system or negentropy term, S_1 is the final entropy of the system with the information. Brillouin (1953) showed that whenever information is obtained about a physical system there is an increase of entropy in the system or in its surroundings. Therefore, a generalized definition of thermodynamic entropy was proposed to be the difference of thermodynamical entropy of the system (ΔS) minus the information (I) possessed by an external observer about the system (Brillouin, 1953; Plenio & Vitelli, 2001).

Karnani et al. (2009) set out to show that information is physical, and that the 2nd law of thermodynamics, although customarily formulated as the principle of increasing disorder in the world, it actually dictates a universal tendency of creating hierarchical structures that develop towards better structures for energy dispersal (or, equivalently, for the increase of entropy). Starting from a generally accepted principle that all forms of information processing, i.e. observing, encoding, transmission, decoding, are natural processes governed by the laws of physics (Brillouin, 1953; Sharma & Annila, 2007; Prigogine, 1978), Karnani et al. (2009) show mathematically that thermodynamic entropy can be used as a measure for information and that it is impossible to create or destroy information without a change in free energy. The rate of change in thermodynamical entropy, under the 2nd law of thermodynamics, was shown to explain the emergence of natural hierarchical structures in the organization of information and of communication systems (Annila & Kuismanen, 2009; Hartonen & Annila, 2012; Karnani et al., 2009): energy disperses according to the principle of maximum entropy production, which means that energy is most efficiently dispersed when the largest flows move from high density to low density to increase entropy most rapidly. The authors conclude that it is rather the thermodynamics principles and thermodynamic entropy that explain information processing/transmission and communication, than information theory (Karnani et al., 2009).

According to this view, brain activity and cognitive processes, as a reflection of brain activity, were hypothesized to be governed by the quest to consume free energy in the least time possible, which means a drive towards increasing entropy dissipation, in accord with the 2nd law of thermodynamics (Varpula et al., 2013).

In other words, brain activity and cognitive processes are not driven by an entropy-reduction principle, but by a principle of increasing entropy dissipation, which is a natural result of the quest to consume free energy in the least time. Taking this view into account, we propose that a specific information-encoding cognitive mechanism such as (linguistic or general) rule induction is a natural free-energy consuming mechanism that generates increasing thermodynamic entropy in the environment, under the rule of the 2nd law of thermodynamics.

1.4 Natural tendency towards structure and the laws of thermodynamics

After having reformulated the 2nd law of thermodynamics and the concept of entropy so that the misconception of flow from order towards disorder is removed from the discussion, the obvious tendency of life and nature towards structure can be understood under the laws of thermodynamics. In fact, one thermodynamic view has proposed that hierarchical structure in nature emerges as a result of the 2nd law of thermodynamics: the quest to consume free energy in the least time leads to structures that are better and faster at doing it (Annila & Kuismanen, 2009; Annila & Annila, 2008; Varpula et al., 2013). Thus, the spontaneous organization into structures tends towards functional complexity in order to reach high entropy (Annila & Annila, 2008). As one of the numerous spontaneous structures, the neural networks of the brain have developed, like any other natural networks, as pathways to facilitate energy transduction, i.e. transformation into other forms of energy, from sensory signals to neural signals (Hartonen & Annila, 2012; Varpula et al., 2013). Hence, the view of cognition under the governance of the laws of thermodynamics takes concrete shape: information as a flow of energy enters brain's neural network, as in the case of any open system interacting with its surroundings, and it gets transmitted through the pathways of energy transduction, by consuming the free energy associated with the physical representations of the signals (Varpula et al., 2013). The flow of energy searches for those structures of pathways or patterns of energy transduction that allow for faster spread of energy, thus increasing entropy dissipation, as per the 2nd law (Annila & Annila, 2008; Hartonen & Annila, 2012; Varpula et al., 2013).

Another thermodynamic view has taken one step further and suggested that the 2nd law might not be sufficient to explain the entire design and structures of nature, animate and inanimate systems alike, but it needs to be accompanied by the constructal law (Bejan, 1997; Bejan & Zane, 2012). The constructal law is another law of thermodynamics, another first principle in physics, just like the 2nd law of thermodynamics. According to this law, the world is a flow system made of an immense collection of many flow systems. Everything that moves, animate or inanimate, is a flow system. In short, a flow system can be described as follows: something that flows, a current (e.g. heat, mass, fluid, information), the rate at which the current flows, and the structure that hosts the current and facilitates it (the background).

The movement of the current encounters resistance created by the background, such as friction, which acts as a brake on the engine (the structure) that carries the current. Resistance opposes movement and it creates loss of energy on the way. For example, a heat current must face a temperature difference in order to flow, and an electric current must overcome a difference in potential in order to flow. However, the brakes should not be understood as limitations in the design because they cause loss of energy, in fact they are necessary in the structure, because otherwise the currents will accelerate incessantly until they spin out of control, without being able to create useful work (Bejan & Zane, 2012). For example, a river's basin is the structure (design) that enables the current of water to flow from an area to a point (a river's mouth), while overcoming the resistance of the land. Similar examples of tree-like structures are ubiquitous in nature, and they facilitate the flow of other currents: oxygen through the air passages of lungs, electricity through a lightning bolt, electrical signals through the dendrites of neurons in the brain, etc. Other more complex designs than tree-like structures include animal and human design that evolve in such a way that they move mass more efficiently across the landscape: they develop body mass and shapes with a better ratio of distance covered by unit of useful energy used (Bejan & Zane, 2012).

The constructal law says that the movement enables the emergence in time of very particular designs and structures that facilitate better flow through a resistant landscape. Thus, every flow system in nature did not emerge and evolve randomly, but it is structured according to an engine-and-brake design containing many engine-and-brake systems, e.g. winds, ocean currents, rivers, animals, humans, man-made machinery, science, information, etc. All these systems evolve towards better-flowing engines, on one side (i.e. more efficiency in producing work from the fixed energy input, hence less loss of energy) and towards more effective brakes, on the other side (i.e. more dissipation of energy into the environment, hence higher rates of entropy generation). This design emerges because it facilitates the flow of energy (Bejan & Zane, 2012).

2. Towards a thermodynamic model of rule induction

In biosciences, the memory of biological systems was proposed to have evolved and to have been designed in accordance with the thermodynamic principles: the neural network is an energy transduction system, where the external signals (energy) from the environment enter the perception system as visual, auditory, etc. signals and are transformed in neural signals, which are dispersed as flows of energy, along the neural pathways (Annila, 2016a; Varpula et al., 2013). The neural signals search for and flow through the pathways that facilitate energy dispersal in the least time: neural networks are viewed as yet another natural network designed according to the 2nd law of thermodynamics, as an energy transduction system that has evolved towards efficient least-time consumption of free energy (Annila, 2016; Hartonen & Annila, 2012; Sharma & Annila, 2007).

Natural networks and structure emergence in natural biological systems were proposed to be governed and driven by the quest to consume free energy in the least time (Hartonen & Annila, 2012; Sharma & Annila, 2007). The constructal law (Bejan, 1997, Bejan & Zane, 2012) would predict that the neural signals do not actually "search" for the pathways that facilitate energy dispersal. Rather, information transmission as a flow system (i.e. the propagation of neural signals) creates those pathways into a structure that facilitates better energy dispersal (just like a river creates its basin to facilitate its flow). Thus, the self-organization of the neural network, just like any other natural (biological) network, is governed by the laws of thermodynamics.

In accord with these views, we propose that rule induction – as an information encoding mechanism that draws on memory resources – can be envisaged as a reflection, at the cognitive level, of a constructal design of neural networks or self-organization, that evolved for the purpose of facilitating better information flow. Thus, rule induction can be formulated according to a thermodynamics framework.

Similarly, regarding another information-theoretic approach to linguistics – Zipf's law – Bejan & Zane (2012) suggest that the constructal law predicts Zipf's empirical law (Zipf, 1949). This is a word frequency law thought to be one of the few truly universal in language (Montemurro & Zanette, 2011; van Egmond, 2018; for a relevant review, see Piantadosi, 2014). The prediction made by Bejan & Zane (2012) is based on the claim that the Zipfian log-log frequency-rank distribution found in language can be directly derived from the mathematical formulation of the constructal law (just like the log-log size-to-rank distribution of settlements in Europe from 1600 to 1980 – Bejan & Zane, 2012). Explained constructally, for the purpose of facilitating the flow of information, in written and spoken communication, the frequency-rank distribution of words creates the hierarchy of channels that evolved under the constructal law: few large channels (the most frequent words – e.g. "the", "to", "of") and many small channels (low frequency words – e.g. "egregious", "ameliorate") (Bejan & Zane, 2012). We suggest future research should investigate the derivation of the Zipfian word frequency distribution specific to language from the mathematical formulation of the constructal law (Bejan & Lorente, 2010), which brings a more refined and biophysically plausible alternative to the Zipfian distribution.

Based on the views and evidence revealed in this review, we propose an innovative thermodynamic model for rule induction. We propose that rule induction as an information encoding mechanism is a reflection at the cognitive level of an evolutionary consequence that follows from the laws of thermodynamics: evolving towards structures (or rules) that facilitate more efficient energy transduction – flow of information (Bejan, 1997; Bejan & Zane, 2012), that is consuming more free energy in the least time, with an increase in entropy as a consequence. Our proposal is not only in accord with recent theoretical advances in physics and biophysics, as reviewed above (Annila & Kuismanen, 2009; Annila & Annila, 2008; Bejan, 1977; Varpula et al., 2013), but also consistent with recent neuroscientific findings (Guevara Erra, Mateos, Wennberg, & Velazquez, 2016; McIntosh, Kovacevic, & Itier, 2008; Protzner, Valiante, Kovacevic, McCormick, & McAndrews, 2010; Velazquez et al., 2019). According to these findings the healthy brain tends towards increased entropy as opposed to unhealthy brain states, which are characterized by lower levels of brain signal entropy.

Specifically, we propose that rule induction is a cognitive mechanism whose purpose is to create structure by the same design principles dictated by the constructal law, as all the other natural structure developing phenomena (Bejan & Zane, 2012; Bejan & Lorente, 2010), in order to facilitate the flow of information (i.e. to consume free energy in the least time). Facilitation of free energy consumption increases the entropy in the world, as per the 2nd law of thermodynamics. In other words, rule induction happens as a result of the brain's tendency to consume more free energy more efficiently (i.e. with the least loss and in the least time), and consequently, to dissipate more entropy into the environment, under the governance of the 2nd law of thermodynamics.

How does rule induction facilitate more efficient consumption of free energy in the brain? As it was theorized and shown by all the studies in this dissertation, from an information-theoretic point of view based on Shannon's noisy-channel coding theory (Shannon, 1948), rule induction is a phased mechanism driven by an increase of input entropy within the bounds set by the finite channel capacity, i.e. the maximum rate of information transmission (bits/s). This enables the gradual transition from high-specificity *item-bound encoding* to high-generality *category-based encoding*. Based on these information-theoretic concepts, our entropy and noisy-channel model for rule induction posits that the change in encoding method, i.e. from a high-specificity *item-bound encoding* to a high-generality *category-based encoding*, is driven by a kind of a regulatory mechanism. This regulatory mechanism moves from an inefficient encoding method (with loss of information), to a better, more efficient encoding method, which allows for higher input entropy to be transmitted reliably (with the least loss of information) and faster (at the maximum rate of information per second, i.e. at *channel capacity*). Reliability of encoding should be understood as given by the least loss of information (caused by noise interference).

However, as we pointed out in Chapter 6 (Radulescu, Murali, Wijnen, & Avrutin, 2021), the information-theoretic model and Shannon's noisy-channel coding theory alone cannot explain *why* and *how* the change happens from a high-specificity *item-bound encoding* to a high-generality *category-based encoding.* Shannon's channel capacity theory posits that an increase of the rate of equivocation (i.e. loss of information) renders the encoding method inefficient, and that "it is possible to find another encoding method", but it does not specify *how* this encoding emerges and how it is designed. Also Shannon's coding theory, and an information-theoretic model based on it do not offer a direct explanation as to what *drives the need* to find another encoding method for our biological encoding system. In other words, *why* does rule induction happen? Hence, in Chapter 6 we proposed an extension of our information-theoretic model by linking it with the dynamic systems hypothesis which applies to self-organizing systems. In self-organization, entropy and noise are a driving force towards new structures (Prigogine & Stengers, 1984; Schneider & Sagan, 2005). Here, we further propose for future research a joint information-theoretic and thermodynamic model based on the constructal law, which accounts for and predicts self-organization (Bejan & Zane, 2012; Bejan & Lorente, 2010).

We suggest that this joint model could offer an answer to the questions *why* and *how* rule induction – with its two flavors: *item-bound* and *category-based generalization* – happens. More specifically, under the thermodynamic framework proposed in this chapter, rule induction is hypothesized to be an information processing and encoding mechanism by which structures (rules) emerge to allow better and faster flow of energy (in the form of information). Better flow is meant in the sense of more efficient flow of energy or efficient consumption of free energy, and faster flow in the sense of consumption of free energy in the least time possible. This proposal is in line with the information-theoretic hypotheses of the model, as summarized in the previous paragraph:

the regulatory mechanism shapes the inefficient *item-bound encoding* (with loss of information) that can only allow for low entropy to flow (to be transmitted), into a better, more efficient *category-based encoding,* that allows higher entropy to flow at a faster rate, i.e. maximum rate of information transmission (*channel capacity*). In other words, the form of encoding transitions from a form that allows for less information to flow – or less free energy transduction (less entropy dissipation), to a form of encoding that allows for more information to flow, i.e. with less loss of information – or more efficient free energy consumption, in order to facilitate higher entropy dissipation.

How can we envisage *category-based encoding* as a form of encoding that facilitates more efficient consumption of free energy (information) than *item-bound encoding*? As we argued in Chapter 5 (Radulescu, Kotsolakou, Wijnen, Avrutin & Grama, 2021), when the inflow of information increases (i.e. high input entropy per second), since the *channel capacity* cannot be exceeded, this calls for a more efficient encoding method such that the actual rate of transmission achieves its maximum, to match the *channel capacity*. Specifically, when the source entropy per second is higher than the available *channel capacity*, the high-specificity *item-bound generalization* becomes inefficient and prone to many errors. Therefore, the information cannot be encoded with a high-fidelity method (i.e. probability matching to the input), because this encoding method gives rise to high loss of information (i.e. increased rate of equivocation).

Thus, the excess of entropy entering the channel results into erasing bindings between items, and reorganizing the redundant (shared) and non-redundant (specific) features of items in order to erase or "forget" insignificant features. This leads to re-grouping the items in categories, that would allow for an infinite number of other novel items to be processed and encoded in the category, i.e. it would allow for better and faster flow of information with higher entropy. As a result, by yielding a more general (less specific) *category-based encoding*, higher *input entropy* can be processed and encoded at the same *channel capacity,* with less loss of information, so more efficiently. In thermodynamic terms, the higher amount of information (higher entropy) means larger amounts of energy are consumed, consequently higher entropy is dispersed (as predicted by the quest to consume free energy in the least time – Hartonen & Annila, 2012; Sharma & Annila, 2007). Also, regrouping into categories calls for higher rates of information erasure (of specific non-shared features), which entails higher energy consumption, thus higher entropy dissipation, as per Landauer's principle (Landauer, 1961).

Furthermore, noise creates loss of information just like imperfections ("brakes") in thermodynamic systems cause heat leaks (Bejan & Zane, 2012; Bejan & Lorente, 2010). According to the constructal law, the "brakes" in the engine-and-brake configuration of any flow system are not a limitation, but they are necessary. In accord with the constructal law, we suggest that noise is necessary in rule induction. As we hypothesized in Chapter 6, noise creates loss of information, which leads to the need for re-structuring the information in such a way that information is more efficiently transmitted, i.e. with the least loss of information. This hypothesis is in line with Shannon's *noisy-channel coding* *theory* (Shannon, 1948) as we showed in Chapter 6, but is not directly predicted or explained by it. However, it is predicted by the constructal law, under our proposed thermodynamic model for rule induction. This hypothesis is supported by recent findings in neurosciences (discussed below), which show a healthy brain is a 'noisy brain' (McIntosh et al., 2008; Protzner et al., 2010).

Further, this model predicts the design of rule induction – *how* the encoding forms are designed – according to the constructal design of a few large channels of energy dispersal and many small channels (Bejan & Zane, 2012; Lorente & Bejan, 2010): a few large channels for information flow – few categories (via *category-based generalization)* and many small channels – many specific items and item-bound relations (*item-bound generalization*). This configuration (design) emerges because this is the most efficient way for information to flow, i.e. to be transmitted. Further research should probe whether the distribution of this hierarchy of channels (few grammatical categories and many lexical/semantic specific items) directly follows from the mathematical formulation of the constructal law (Bejan & Lorente, 2004; 2010). Preliminary evidence from recent proposals on unzipping the Zipf's law in language (Lestrade, 2017) seems to point into this direction, namely that only when considering an interaction of syntactic criteria (grammatical categories) and semantic criteria (item specificity within the class) does Zipf's law hold, with its frequency-rank distribution and line curvature. As mentioned above, Zipf's law could be predicted by the constructal law (as per the proposal of the author of the constructal law himself – Bejan & Zane, 2012). If proven to be the case, the elusive origins of Zipf's law together with the mechanism of rule induction in language might be finally unveiled.

In conclusion, according to the 2nd law of thermodynamics, rule induction happens as a natural result of the tendency of the brain's neural networks (and, as a reflection, the cognitive system) to consume free energy (in the form of information) in the least time possible, which in turn increases the entropy dissipation. The constructal law predicts the generation of the particular hierarchical structure (items and categories) that facilitates the efficient information transmission, as a flow of energy.

While we have proposed some theoretical grounds for a thermodynamic framework to account for rule induction, it remains a more intricate and challenging task to devise a practical and experimental framework to test this theoretic framework/model. Notably, some studies used a method proposed by Costa, Goldberger, & Peng (2005) – Multiscale Entropy Estimation of temporal signal complexity[37] – in order to estimate the complexity of the physiological signal measured with EEG (McIntosh et al., 2008) and iEEG (Protzner et al., 2010). Also, a handful of recent studies focusing on a thermodynamic view on the emergence of cognition and consciousness (Guevara Erra et al., 2016;

[37] Explained briefly at https://sapienlabs.org/understanding-multiscale-entropy/ and at https://archive.physionet.org/physiotools/mse/tutorial/tutorial.pdf

Velazquez et al., 2019) made a step forward in this sense, by making a concrete proposal on an estimation of the Gibs free energy and the dispersal of entropy content in the brain, by using physiological signals from MEG, iEEG and scalp EEG recordings.

Although the scope of these studies and their methodologies are different, they converge on similar hypotheses and findings about the brain and its energy consumption and entropy. They propose methods of estimating (quantifying) energy in the brain in probabilistic terms with direct applications to neuroscientific research (Velazquez et al., 2019). Guevarra Erra et al. (2016) found increasing entropy in the brain in conscious awareness states, which is in accord with thermodynamics-based proposals that the brain, as a thermodynamic system, naturally tends towards better and faster consumption of free energy, hence increasing entropy dissipation (Annila, 2016). Conversely, Guevarra et al. (2016) found lower entropy to be a characteristic of the unconscious or altered states of alertness (eyes closed).

Another line of research that we consider to bring preliminary evidence for our proposed thermodynamic model are the studies by Protzner et al. (2010) and McIntosh et al. (2008). Protzner et al. (2010) recorded iEEG from the hippocampi of participants while they were performing a memory task, and found higher signal entropy (measured using Multiscale Entropy Estimation) in the healthy hippocampus as compared to the epileptogenic hippocampus. They concluded that brain signal entropy is a biomarker of neuronal system integrity.

McIntosh et al. (2008) measured brain signal variability, in children and adults, using EEG (and quantified entropy with the Multiscale Entropy Estimation method), while participants were performing a face recognition task. They found that brain signal entropy increases in children from 8 to 15 years old, and even to a higher extent in adults. Further, when compared with participants' performance on the face recognition task, results showed that higher brain signal entropy correlates with reduced behavioral inconsistency in the task performance, and thus with better performance at information processing. Authors concluded that brain signal entropy or in other words "brain noise" is a marker of healthy brain functioning – a noisy brain is a healthy brain (McIntosh et al., 2008).

Therefore, we propose that further research into the rule induction phenomenon should employ EEG methods to measure the physiological response of the brain as a composite energy consumption signature, while engaged in rule induction tasks, such as the artificial grammar tasks employed by the studies of this dissertation. Either the Multiscale Entropy Estimation method (McIntosh et al., 2008; Protzner et al., 2010) or the method proposed by Guevara Erra et al. (2016) and further developed in Velazquez et al. (2019), or both methods simultaneously could be used. The aim should be to obtain estimations of entropy and energy dispersal in the relevant brain areas that can be associated with rule induction, and more specifically with a transition from *item-bound encoding* to *category-based encoding*. The research question would be to probe whether brain signal entropy increases as a function of increasing input entropy, in line with the predictions of our proposed information-theoretic and thermodynamic model. And if such an increase in brain signal entropy correlates positively with the behavioral tendency to move from *item-bound* to *category-based generalization.*

Such experiments designed to test our joint information-theoretic and thermodynamic model of rule induction would be insightful for shedding light on previous entropy-based hypotheses on language learning and processing. The entropy-reduction hypothesis was proposed – as an essential survival mechanism of open systems – to underlie self-organization in many studies in biology (Nicolis & Prigogine, 1989; Prigogine, 1978). Entropy reduction was also proposed as a trademark of self-organization into new and better structures for problem solving in cognitive sciences (Stephen et al., 2009). Similarly, in information-theoretic psycholinguistic studies, entropy reduction was proposed as the underlying principle for language comprehension and sentence processing (Hale, 2006; Levy, 2008; Linzen & Jaeger, 2016; Venhuizen et al., 2019a; 2019b), and also for rule learning and regularization (Ferdinand et al., 2019). The findings of EEG experiments with a thermodynamic interpretation, under the model we propose here, would shed some light on what might look like an apparent contradiction, or at least a slight glitch in the governance of the 2nd law of thermodynamics. Does the cognitive system (as a reflection of a natural open living system) strive towards a reduction of entropy in order to preserve its internal structure and functionality (Schrödinger, 1944; Prigogine, 1978)? Or does the cognitive system develop structure as a result of striving towards more consumption of free energy, thus as a tendency *towards* entropy, not *against* it (Annila, 2016; Sharma & Anilla, 2007)?

If the findings support the latter view, as predicted by our proposed thermodynamic model, there would be no need to employ ad-hoc assumptions about reducing entropy, or about increasing entropy somewhere in order to reduce it elsewhere (Prigogine, 1978). The same principle of increasing entropy as per the 2nd law of thermodynamics, which is valid for any chemical reaction, would be shown to apply to life processes and, therefore, to neural networks and cognition. A similar view in thermodynamics (Bejan, 2007; 2017; Reis, 2014; 2016) holds the “ad-hoc principles” of reducing and/or maximizing entropy as being divisive, and making sense only if taken together with the constructal law, which unifies these approaches and adds to the 2nd law’s “one way” *irreversible* flow, the tendency of nature to generate a certain kind of structures (design) to facilitate that flow (Bejan & Zane, 2012).

In conclusion, the joint information-theoretic and thermodynamic model proposed here, together with measures of the thermodynamic entropy of the physiological signal associated with rule induction tasks, might help unify apparently opposing hypotheses of entropy-reduction in cognitive sciences and psycholinguistics, with principles of nature’s tendency towards increasing entropy from physics and biosciences, as well as with recent neuroscientific findings of high entropy brain signal as a trademark of healthy brain states.

Our proposed joint information-theoretic and thermodynamic model for estimations of brain information content and energy consumption lays the foundation for a thermodynamics theory on rule induction. This proposal aims at offering a framework and a model to inspire a future research direction based on estimating the thermodynamic costs of information processing, erasing and encoding. We hope that the findings of this research direction will show that rule induction is an encoding mechanism that happens automatically as a direct effect of the laws of thermodynamics, just like photosynthesis happens to a flower in the sunlight. And this is not a metaphor.

References

Abboub, N., Nazzi, T., & Gervain, J. (2016). Prosodic grouping at birth. *Brain and language*, *162*, 46–59. https://doi.org/10.1016/j.bandl.2016.08.002

Abraham, W. C., & Robins, A. (2005). Memory retention--the synaptic stability versus plasticity dilemma. *Trends in neurosciences*, *28*(2), 73–78. https://doi.org/10.1016/j.tins.2004.12.003

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.

Alhama, R. G., & Zuidema, W. (2019). A review of computational models of basic rule learning: The neural-symbolic debate and beyond. *Psychonomic bulletin & review*, *26*(4), 1174–1194. https://doi.org/10.3758/s13423-019-01602-z

Altmann, G. T. (2002). Learning and development in neural networks—the importance of prior experience. Cognition, 85(2), B43–B50.

Altmann, G. T., & Dienes, Z. (1999). Rule learning by seven-month old infants and neural networks. Science, 284(5416), 875–875.

Anderson, J. R. (1993). *Rules of the Mind*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working memory: activation limitations on retrieval. *Cognitive psychology*, *30*(3), 221–256. https://doi.org/10.1006/cogp.1996.0007

Annila, A. (2016a). On the character of consciousness. *Front. Syst. Neurosci. 10*:27. doi: 10.3389/fnsys.2016.00027

Annila, A. (2016b). Natural thermodynamics. *Physica A 444*, 843–852. https://doi.org/10.1016/j.physa.2015.10.105

Annila, A., & Annila, E. (2008). Why did life emerge? *International Journal of Astrobiology, 7*(3-4), 293-300. doi:10.1017/S1473550408004308

Annila, A. & Baverstock, K. (2016). Discourse on order vs. disorder, *Communicative & Integrative Biology, 9*:4, DOI: 10.1080/19420889.2016.1187348

Annila, A., & Kuismanen, E. (2009). Natural hierarchy emerges from energy dispersal. *Bio Systems*, *95*(3), 227–233. https://doi.org/10.1016/j.biosystems.2008.10.008

Aslin, R. N., & Newport, E. L. (2012). Statistical learning: From acquiring specific items to forming general rules. *Current directions in psychological science*, *21*(3), 170–176. https://doi.org/10.1177/0963721412436806

Aslin, R. N., & Newport, E. L. (2014). Distributional Language Learning: Mechanisms and Models of ategory Formation. *Language learning*, *64*(Suppl 2), 86–105. https://doi.org/10.1111/lang.12074

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., Shukla, M., & Emberson, L. L. (2015). Hemodynamic correlates of cognition in human infants. *Annual review of psychology*, *66*, 349–379. https://doi.org/10.1146/annurev-psych-010213-115108

Avery, J.S. (2012). *Information, Theory and Evolution*. London, UK: World Scientific.

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

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

Baddeley, A. (2000). The episodic buffer: a new component of working memory?. *Trends in cognitive sciences*, *4*(11), 417–423. https://doi.org/10.1016/s1364-6613(00)01538-2

Baddeley, A. (2007). *Oxford psychology series: Vol. 45. Working memory, thought, and action.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198528012.001.0001

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63,* 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2015). *Memory* (2nd ed.). Psychology Press.

Bejan, A. (1996). *Entropy Generation Minimization*, CRC Press, Boca Raton, FL

Bejan, A. (1997a). Constructal-theory network of conducting paths for cooling a heat generating volume, *Int. J. Heat Mass Trans., 40*, 799–816

Bejan, A. (1997b). Theory of organization in Nature: Pulsating physiological processes, *Int. J. Heat Mass Trans., 40*, 2097–2104

Bejan, A. (1997c). Constructal tree network for fluid flow between a finite-size volume and one source or sink, *Int. J. Thermal Sci., 36*, 592–604

Bejan, A. (1997d). *Advanced Engineering Thermodynamics*, 2nd ed., Wiley, New York

Bejan, A. (2017) Evolution in thermodynamics. *Applied Physics Reviews 4*, 011305; doi: 10.1063/1.4978611

Bejan, A., & Lorente, S. (2004). The constructal law and the thermodynamics of flow systems with configuration, *Int. J. Heat Mass Trans., 47*, 3203–3214

Bejan, A., & Lorente, S. (2010). The constructal law of design and evolution in nature. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *365*(1545), 1335–1347. https://doi.org/10.1098/rstb.2009.0302

Bejan, A., & Marden, J. H. (2009). The constructal unification of biological and geophysical design. *Physics of life reviews*, *6*(2), 85–102. https://doi.org/10.1016/j.plrev.2008.12.002

Bejan, A., & Zane, J. P. (2012). *Design in Nature: How the Constructal Law Governs Evolution in Biology, Physics, Technology, and Social Organization*. Doubleday Books.

Bennett C. H. (1982). The thermodynamics of computation—a review. *Int. J. Theor. Phys*. 21, 905–940. 10.1007/BF02084158

Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(9), 3253–3258. https://doi.org/10.1073/pnas.1113380109

Bérut, A., Arakelyan, A., Petrosyan, A., Ciliberto, S., Dillenschneider, R., and Lutz, E. (2012). Experimental verification of Landauer's principle linking information and thermodynamics. Nature 483, 187–189. doi: 10.1038/nature10872.

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

Boersma, P. and Weenink, D. (2014). Praat: doing phonetics by computer [Computer program]. Version 5.4.02, retrieved in 2014 from http://www.praat.org/

Boersma, P., & Weenink, D. (2019). Praat: Doing phonetics by computer [Computer program]. Version 6.0.49, Retrieved from http://www.praat.org/

Boltzmann L. (1877). Uber die beziehung dem zweiten Haubtsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive den Sätzen über das Wärmegleichgewicht [On the relationship between the second main theorem of mechanical heat theory and the probability calculation with respect to the results about the heat equilibrium]. *Wien. Ber*. 76, 373–435.

Bouchon, C., Nazzi, T., & Gervain, J. (2015). Hemispheric Asymmetries in Repetition Enhancement and Suppression Effects in the Newborn Brain. *PloS one*, *10*(10), e0140160. https://doi.org/10.1371/journal.pone.0140160

Brillouin L. (1953). The negentropy principle of information. *J. Appl. Phys*. 24, 1152–1163. 10.1063/1.1721463

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.

Burgoyne, A. P., Hambrick, D. Z., & Altmann, E. M. (2019). Is working memory capacity a causal factor in fluid intelligence?. *Psychonomic bulletin & review*, *26*(4), 1333–1339. https://doi.org/10.3758/s13423-019-01606-9

Candice C. Morey & Jonathan T. Mall (2012) Cross-domain interference costs during concurrent verbal and spatial serial memory tasks are asymmetric, The Quarterly Journal of Experimental Psychology, 65:9, 1777-1797, DOI: 10.1080/17470218.2012.668555

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*(3), 404–431. https://doi.org/10.1037/0033-295X.97.3.404

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Chambers, K., Onishi, K., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition, 87*, B69-B77.

Christiansen, M. H., & Arnon, I. (2017). More Than Words: The Role of Multiword Sequences in Language Learning and Use. *Topics in cognitive science*, *9*(3), 542–551. https://doi.org/10.1111/tops.12274

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *The Behavioral and brain sciences*, *31*(5), 489–558. https://doi.org/10.1017/S0140525X08004998

Christiansen, M., Conway, C., & Curtin, S. (2000). A connectionist single mechanism account of rule-like behavior in infancy. In *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*, (pp. 83–88)

Christiansen, M.H. & Curtin, S.L. (1999). The power of statistical learning: No need for algebraic rules. In *The Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 114-119). Mahwah, NJ: Lawrence Erlbaum.

Chubala, C. M., & Jamieson, R. K. (2013). Recoding and representation in artificial grammar learning. *Behavior research methods*, *45*(2), 470–479. https://doi.org/10.3758/s13428-012-0253-6

Cocchini, G., Logie, R. H., Della Sala, S., MacPherson, S. E., & Baddeley, A. D. (2002). Concurrent performance of two memory tasks: Evidence for domain-specific working memory systems. Memory & Cognition, 30, 1086–1095.

Collell, G., & Fauquet, J. (2015). Brain activity and cognition: a connection from thermodynamics and information theory. *Frontiers in psychology*, *6*, 818. https://doi.org/10.3389/fpsyg.2015.00818

Colombo, M., & Wright, C. (2018). First principles in the life sciences: the free-energy principle, organicism, and mechanism. *Synthese* . https://doi.org/10.1007/s11229-018-01932-w

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30*(2), 163–184. https://doi.org/10.1016/S0160-2896(01)00096-4

Conway, A. R., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: the importance of working memory capacity. *Psychonomic bulletin & review*, *8*(2), 331–335. https://doi.org/10.3758/bf03196169

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic bulletin & review*, *12*(5), 769–786. https://doi.org/10.3758/bf03196772

Conway, A. R., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in cognitive sciences*, *7*(12), 547–552. https://doi.org/10.1016/j.tics.2003.10.005

Costa, M., Goldberger, A. L., & Peng, C. K. (2005). Multiscale entropy analysis of biological signals. *Physical review. E, Statistical, nonlinear, and soft matter physics*, *71*(2 Pt 1), 021906. https://doi.org/10.1103/PhysRevE.71.021906

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin, 104*(2), 163–191. https://doi.org/10.1037/0033-2909.104.2.163

Cowan, N. (1995). *Oxford psychology series, No. 26.Attention and memory: An integrated framework.* Oxford University Press.

Cowan, N. (1997). *The development of working memory.* In N. Cowan (Ed.), *Studies in developmental psychology. The development of memory in childhood* (p. 163–199). Psychology Press/Erlbaum (UK) Taylor & Francis.

Cowan, N. (1999). *An Embedded-Processes Model of working memory.* In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (p. 62–101). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.006

Cowan, N. (2005). *Essays in cognitive psychology. Working memory capacity.* Psychology Press. https://doi.org/10.4324/9780203342398

Cowan, N., Nugent, L. D., Elliott, E. M., Ponomarev, I., & Saults, J. S. (1999). The role of attention in the development of short-term memory: age differences in the verbal span of apprehension. *Child development, 70*(5), 1082–1097. https://doi.org/10.1111/1467-8624.00080

Cowan, N., Nugent, L. D., Elliott, E. M., & Saults, J. S. (2000). Persistence of memory for ignored lists of digits: areas of developmental constancy and change. *Journal of experimental child psychology, 76*(2), 151–172. https://doi.org/10.1006/jecp.1999.2546

Cox, G. E., & Shiffrin, R. M. (2017). A dynamic approach to recognition memory. *Psychological review*, *124*(6), 795–860. https://doi.org/10.1037/rev0000076

DeCastro, A. (2013). The thermodynamic cost of fast thought. *Minds Mach. 23*, 473–487. doi:10.1007/s11023-013-9302-x

Dehn, M. J. (2017). How working memory enables fluid reasoning. Applied Neuropsychology: Child, 6(3), 245-247.doi:10.1080/21622965.2017.1317490

Del Castillo L. F., & Vera-Cruz P. (2011). Thermodynamic formulation of living systems and their evolution. *J. Mod. Phys.* 2, 379–391. 10.4236/jmp.2011.25047

Dijksterhuis, A., & Nordgren, L. F. (2006). A Theory of Unconscious Thought. *Perspectives on Psychological Science*, *1*(2), 95–109. https://doi.org/10.1111/j.1745-6916.2006.00007.x

Dominey, P. F., & Ramus, F. (2000). Neural network processing of natural language I. sensitivity to serial, temporal and abstract structure of language in the infant. Language and Cognitive Processes, 15(1), 87–127.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & van der Vreken, O. (1996) The MBROLA project: towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In *Proceedings of the fourth international conference on spoken language processing, 96 (3)*: 1393–1396.

Emberson, L. L., Conway, C. M., & Christiansen, M. H. (2011). Timing is everything: changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. *Quarterly journal of experimental psychology (2006)*, *64*(5), 1021–1040. https://doi.org/10.1080/17470218.2010.538972

Endress, A. D. (2013). Bayesian learning and the psychology of rule induction. Cognition, 127(2), 159–176.

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. Cognition, 105, 247–299.

Endress, A. D., & Bonatti, L. L. (2016). Words, rules, and mechanisms of language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 7*(1), 19–35.

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

Endress, A. D., Nespor, M. & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences 13*, 348–53.

Endress, A. D., Scholl, B. J., & Mehler, J. (2005). The Role of Salience in the Extraction of Algebraic Rules. *Journal of Experimental Psychology: General, 134*(3), 406–419. https://doi.org/10.1037/0096-3445.134.3.406

England, J.L. (2013). Statistical Physics of self-replication*. J Chem Phys 139*:121923; PMID:24089735; http://dx.doi.org/10.1063/1.4818538

England, J.L. (2015). Dissipative Adaptation in Driven Self-assembly. *Nature Nanotech 10*:920; PMID:26530021; http://dx.doi.org/10.1038/nnano.2015.250

Ferdinand, V. (2015). Inductive evolution: Cognition, culture, and regularity in language. PhD thesis, University of Edinburgh

Ferdinand, V., Kirby, S., & Smith, K. (2019). The cognitive roots of regularization in language. *Cognition*, *184*, 53–68. https://doi.org/10.1016/j.cognition.2018.12.002

Feynman R. P., Leighton R. B., Sands M., Hafner E. M. (1965). The feynman lectures on physics. *Am. J. Phys*. 33, 750–752. 10.1119/1.1972241

Fillmore, P. T., Richards, J. E., Phillips-Meek, M. C., Cryer, A., & Stevens, M. (2015). Stereotaxic Magnetic Resonance Imaging Brain Atlases for Infants from 3 to 12 Months. *Developmental neuroscience*, *37*(6), 515–532. https://doi.org/10.1159/000438749

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

Frankland, P.W., Köhler, S., and Josselyn, S.A. (2013). Hippocampal neurogenesis and forgetting. *Trends Neurosci. 36*, 497–503.

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13*, 293–301.

Friston, K. (2010). The free-energy principle: A unified brain theory? Nature *Reviews Neuroscience, 11*, 127–138.

Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese, 159(3)*, 417–458. https://doi.org/10.1007/s11229-007-9237-y

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

Gartman, L. M., & Johnson, N. F. (1972). Massed versus distributed repetition of homographs: A test of the differential-encoding hypothesis. *Journal of Verbal Learning & Verbal Behavior, 11*(6), 801–808. https://doi.org/10.1016/S0022-5371(72)80016-1

Gasser, M., & Colunga, E. (2000). Babies, variables, and relational correlations. In *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*, (Vol. 22, p. 160): Lawrence Erlbaum Associates.

Gathercole, S.E. (1998), The Development of Memory. *Journal of Child Psychology and Psychiatry, 39*: 3-27. https://doi.org/10.1111/1469-7610.00301

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

Gerken, L.A. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition, 115(2)*,362–366.

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

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.

Gerken, L., & Quam, C. (2017). Infant learning is influenced by local spurious generalizations. *Developmental science*, *20*(3), 10.1111/desc.12410. https://doi.org/10.1111/desc.12410

Gervain, J., Berent, I., & Werker, J. F. (2012). Binding at birth: The newborn brain detects identity relations and sequential position in speech. *Journal of Cognitive Neuroscience, 24(3)*, 564–574.

Gervain, J., Macagno, F., Cogoi, S., Peña, M., & Mehler, J. (2008). The neonate brain detects speech structure. *Proceedings of the National Academy of Sciences of the United States of America, 105(37)*, 14222–14227.

Gervain, J., & Werker, J. F. (2012). Learning non-adjacent regularities at age 0 ; 7. *Journal of Child Language, FirstView,* 1–13.

Giurfa, M., Zhang, S., Jenett, A., Menzel, R., & Srinivasan, M. V. (2001). The concepts of 'sameness' and 'difference' in an insect. *Nature*, *410*(6831), 930–933. https://doi.org/10.1038/35073582

Godden, D., & Baddeley, A. (1980). When does context influence recognition memory? *British Journal of Psychology, 71*(1), 99–104. https://doi.org/10.1111/j.2044-8295.1980.tb02735.x

Gómez, R.L. (2002). Variability and detection of invariant structure. *Psychological Science, 13(5),* 431 – 436.

Gómez, R.L., Bootzin, R. R., & Nadel, L. (2006). Naps promote abstraction in language-learning infants. *Psychological science*, *17*(8), 670–674. https://doi.org/10.1111/j.1467-9280.2006.01764.x

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

Gómez, R., & Maye, J. (2005). The Developmental Trajectory of Nonadjacent Dependency Learning. *Infancy, 7*(2), 183–206. https://doi.org/10.1207/s15327078in0702_4

Grama, I. C., Kerkhoff, A., & Wijnen, F. (2016). Gleaning structure from sound: The role of prosodic contrast in learning non-adjacent dependencies. *Journal of Psycholinguistic Research, 45*(6), 1427–1449. https://doi.org/10.1007/s10936-016-9412-8

Griffiths, T.L., & Tenenbaum, J.B. (2007). From mere coincidences to meaningful discoveries. *Cognition 103*(2), 180–226.

Guevara Erra, R., Mateos, D.M., Wennberg, R., & Velazquez, J.L. (2016). Statistical mechanics of consciousness: Maximization of information content of network is associated with conscious awareness. *Physical Review, E,* 94(5-1), 52402*.*

Hale J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science, 30*(4), 643–672. https://doi.org/10.1207/s15516709cog0000_64

Hardt, O., Nader, K., & Wang, Y. T. (2013). GluA2-dependent AMPA receptor endocytosis and the decay of early and late long-term potentiation: possible mechanisms for forgetting of short- and long-term memories. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *369*(1633), 20130141. https://doi.org/10.1098/rstb.2013.0141

Hardt, M., Recht, B., & Singer, Y. (2016) Train faster, generalize better: Stability of stochastic gradient decent. *Proceedings of the 33rd International Conference on Machine Learning*, New York, USA.

Harley, H. E., Putman, E. A., & Roitblat, H. L. (2003). Bottlenose dolphins perceive object features through echolocation. *Nature*, *424*(6949), 667–669. https://doi.org/10.1038/nature01846

Hartonen, T., & Annila, A. (2012). Natural networks as thermodynamic systems. *Complexity 18*, 53–62. doi:10.1002/cplx.21428

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

Hawkins, D.M. (2004). The problem of overfitting. *J. Chem. Inf. Comput. Sci. 44*, 1–12.

Hinton, G. E., & van Camp, D. (1993). Keeping neural networks simple. In *Proceedings of the International Conference on Artificial Neural Networks*, Amsterdam, pages 11–18. Springer.

Hochmann, J. R., Mody, S., & Carey, S. (2016). Infants' representations of same and different in match- and non-match-to-sample. *Cognitive psychology*, *86*, 87–111. https://doi.org/10.1016/j.cogpsych.2016.01.005

Hudson Kam C. (2019). Reconsidering retrieval effects on adult regularization of inconsistent variation in language. *Language learning and development : the official journal of the Society for Language Development*, *15*(4), 317–337. https://doi.org/10.1080/15475441.2019.1634575

Hudson Kam, C. L. H., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development, 1*(2), 151–195. https://doi.org/10.1207/s15473341lld0102_3

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*(3), 815–821. https://doi.org/10.1037/a0015097

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: when learners change languages. *Cognitive psychology*, *59*(1), 30–66. https://doi.org/10.1016/j.cogpsych.2009.01.001

Jaeger, T. F. (2006) Redundancy and syntactic reduction. Ph.D. thesis, Stanford University.

Jaeger T. F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cognitive psychology*, *61*(1), 23–62. https://doi.org/10.1016/j.cogpsych.2010.02.002

Jamieson, R. K., & Mewhort, D. J. (2010). Applying an exemplar model to the artificial-grammar task: String completion and performance on individual items. *Quarterly journal of experimental psychology (2006)*, *63*(5), 1014–1039. https://doi.org/10.1080/17470210903267417

Kabdebon, C., Leroy, F., Simmonet, H., Perrot, M., Dubois, J., & Dehaene-Lambertz, G. (2014). Anatomical correlations of the international 10-20 sensor placement system in infants. *NeuroImage*, *99*, 342–356. https://doi.org/10.1016/j.neuroimage.2014.05.046

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General, 126*(3), 278–287. https://doi.org/10.1037/0096-3445.126.3.278

Karnani, M., Pääkkönen, K., & Annila, A. (2009) The physical character of information. *Proc. R. Soc. A.***465,** 2155–2175, http://doi.org/10.1098/rspa.2009.0063

Kidd, C., Piantadosi, S.T., & Aslin, R.N. (2012). The Goldilocks Effect: Human Infants Allocate Attention to Visual Sequences That Are Neither Too Simple Nor Too Complex. *PLoS ONE 7*(5): e36399.doi:10.1371/journal.pone.0036399

Kirkaldy J. S. (1965). Thermodynamics of the human brain. *Biophysical journal*, *5*(6), 981–986. https://doi.org/10.1016/S0006-3495(65)86763-7

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.

Koch, K., McLean, J., Segev, R., Freed, M. A., Berry, M. J., 2nd, Balasubramanian, V., & Sterling, P. (2006). How much the eye tells the brain. *Current biology : CB*, *16*(14), 1428–1434. https://doi.org/10.1016/j.cub.2006.05.056

Kumaran, D., Hassabis, D., & McClelland, J.L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends Cogn. Sci. 20*, 512–534.

Kuznetsova, A., Brockhoff, P.B., & Christensen, R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. doi: 10.18637/jss.v082.i13.

La Cerra P. (2003). The first law of psychology is the second law of thermodynamics: the energetic evolutionary model of the mind and the generation of human psychological phenomena. *Hum. Nat. Rev*. 3, 440–447.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science (New York, N.Y.)*, *350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *The Behavioral and brain sciences*, *40*, e253. https://doi.org/10.1017/S0140525X16001837

Landauer R. (1961). Irreversibility and heat generation in the computing process. *IBM J. Res. Dev*. 5, 183–191. 10.1147/rd.53.0183

Landauer R. (1991). Information is physical. *Physics Today 44*, 5, 23. https://doi.org/10.1063/1.881299

Lany, J., & Gómez, R. L. (2008). Twelve-month-old infants benefit from prior experience in statistical learning. *Psychological Science, 19*(12), 1247–1252.

Lany, J., Gómez, R. L., & Gerken, L. (2007). The role of prior experience in language acquisition. *Cognitive Science, 31*, 481–507.

Laughlin, S. B., de Ruyter van Steveninck, R. R., & Anderson, J. C. (1998). The metabolic cost of neural information. *Nature neuroscience*, *1*(1), 36–41. https://doi.org/10.1038/236

Le Bellac, M., Mortessagne, F., & Batrouni, G. (2004). *Equilibrium and Non-Equilibrium Statistical Thermodynamics*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511606571

LeCun, Y. (1989). Generalization and network design strategies. Technical Report CRG-TR-89-4.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Denker, J.S., and Solla, S.A. (1989). Optimal brain damage. In *Advances in Neural Information Processing Systems 2*, R.E. Howard and L.D. Jackel, eds., pp. 598–605.

Lestrade, S. (2017). Unzipping Zipf's law. *PLOS ONE 12*(8): e0181987. https://doi.org/10.1371/journal.pone.0181987

Levy R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Levy, R. & Jaeger, TF. (2007) Speakers optimize information density through syntactic reduction. In: Schlökopf, B.; Platt, J.; Hoffman, T., editors. *Advances in neural information processing systems (NIPS). Vol. 19*, p. 849-856, Cambridge, MA: MIT Press.

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(3), 720–738. https://doi.org/10.1037/a0022639

Linzen, T., & Jaeger, T. F. (2016). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science, 40*(6), 1382–1411. https://doi.org/10.1111/cogs.12274

Little, D. R., Lewandowsky, S., & Craig, S. (2014). Working memory capacity and fluid abilities: The more difficult the item, the more more is better. *Frontiers in Psychology, 5,* Article 239. https://doi.org/10.3389/fpsyg.2014.00239

Little, D. R, Lewandowsky, S., & Griffiths, T. L. (2012). A Bayesian Model of Rule Induction in Raven's Progressive Matrices. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 34.

Lloyd-Fox, S., Blasi, A., & Elwell, C. E. (2010). Illuminating the developing brain: the past, present and future of functional near infrared spectroscopy. *Neuroscience and biobehavioral reviews*, *34*(3), 269–284. https://doi.org/10.1016/j.neubiorev.2009.07.008

Lloyd-Fox, S., Richards, J. E., Blasi, A., Murphy, D. G., Elwell, C. E., & Johnson, M. H. (2014). Coregistering functional near-infrared spectroscopy with underlying cortical areas in infants. *Neurophotonics*, *1*(2), 025006. https://doi.org/10.1117/1.NPh.1.2.025006

Lorente, S., & Bejan, A. (2010) Few large and many small: Hierarchy in movement on earth. *Int. J. Des. Nat. Ecodyn. 5*, 254–267

Lucia, U., Grazzini, G., Montrucchio, B., Grisolia, G., Borchiellini, R., Gervino, G., Castagnoli, C., Ponzetto, A., & Silvagno, F. (2015). Constructal thermodynamics combined with infrared experiments to evaluate temperature differences in cells. *Scientific reports*, *5*, 11587. https://doi.org/10.1038/srep11587

Lutz, E., & Ciliberto, S. (2015). Information: from Maxwell's demon to Landauer's eraser. *Phys. Today 68*(9), 30

Machta, J. (1999) Entropy, information, and computation**.** *American Journal of Physics, 67,* p. 1074

MacKay, D.J.C. (2003). *Information Theory, Inference and Learning Algorithms,* Cambridge University Press.

Marcus, G. (2018). Deep Learning: A Critical Appraisal. *ArXiv, abs/1801.00631*.

Marcus, G. F. (2001). *Learning, development, and conceptual change. The algebraic mind: Integrating connectionism and cognitive science.* The MIT Press.

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

Marr, D. (1982) *Vision: A computational approach*. San Francisco: Freeman and Co.

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.

McIntosh, A. R., Kovacevic, N., & Itier, R. J. (2008). Increased brain signal variability accompanies lower behavioral variability in development. *PLoS computational biology*, *4*(7), e1000106. https://doi.org/10.1371/journal.pcbi.1000106

Meek, J. (2002). Basic principles of optical imaging and application to the study of infant development. *Dev Sci, 5(3)*: 371–380.

Mehta, P., & Schwab, D. J. (2012). Energetic costs of cellular computation. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(44), 17978–17982. https://doi.org/10.1073/pnas.1207814109

Migues, P. V., Liu, L., Archbold, G. E., Einarsson, E. Ö., Wong, J., Bonasia, K., Ko, S. H., Wang, Y. T., & Hardt, O. (2016). Blocking Synaptic Removal of GluA2-Containing AMPA Receptors Prevents the Natural Forgetting of Long-Term Memories. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *36*(12), 3481–3494. https://doi.org/10.1523/JNEUROSCI.3333-15.2016

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

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97. https://doi.org/10.1037/h0043158

Monaghan P. (2017). Canalization of Language Structure From Environmental Constraints: A Computational Model of Word Learning From Multiple Cues. *Topics in cognitive science*, *9*(1), 21–34. https://doi.org/10.1111/tops.12239

Montemurro, M. A., & Zanette, D. H. (2011). Universal entropy of word ordering across linguistic families. *PloS one*, *6*(5), e19875. https://doi.org/10.1371/journal.pone.0019875

Moscovitch, M., Cabeza, R., Winocur, G., & Nadel, L. (2016). Episodic Memory and Beyond: The Hippocampus and Neocortex in Transformation. *Annual review of psychology*, *67*, 105–134. https://doi.org/10.1146/annurev-psych-113011-143733

Mumby D. G. (2001). Perspectives on object-recognition memory following hippocampal damage: lessons from studies in rats. *Behavioural brain research*, *127*(1-2), 159–181. https://doi.org/10.1016/s0166-4328(01)00367-9

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

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

Nicolis, G., & Prigogine, I. (1989). *Exploring Complexity: An Introduction.* W.H. Freeman and Company, New York.

Oberauer, K., & Hein, L. (2012). Attention to information in working memory. *Current Directions in Psychological Science, 21*(3), 164–169. https://doi.org/10.1177/0963721412444727

Oberauer, K., & Lewandowsky, S. (2016). Control of information in working memory: Encoding and removal of distractors in the complex-span paradigm. *Cognition*, *156*, 106–128. https://doi.org/10.1016/j.cognition.2016.08.007

Onnis L., Christiansen M. H., Chater N., & Gómez, P. (2003). Reduction of uncertainty in human sequential learning: evidence from Artificial Grammar Learning, in *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (Mahwah, NJ: Lawrence Erlbaum;), 886–891

Onnis, L., Destrebecqz, A., Christiansen, M.H., Chater, N., Cleeremans, A., (2015). Processing non-adjacent dependencies: A graded, associative account. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages*. John Benjamins.

Onnis, L., Monaghan, P., Christiansen, M. H., & Chater, N. (2004). Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies. In *Proceedings of the 26th annual conference of the cognitive science society*. Mahwah, NJ: Lawrence Erlbaum.

Pacton, S., & Perruchet, P. (2008). An attention-based associative account of adjacent and nonadjacent dependency learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(1), 80–96. https://doi.org/10.1037/0278-7393.34.1.80

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science (New York, N.Y.)*, *298*(5593), 604–607. https://doi.org/10.1126/science.1072901

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

Perez Velazquez, J. L., Mateos, D. M., & Guevara Erra, R. (2019). On a Simple General Principle of Brain Organization. *Frontiers in neuroscience*, *13*, 1106. https://doi.org/10.3389/fnins.2019.01106

Perfors, A. (2012). When do memory limitations lead to regularization? An experimental and computational investigation . *Journal of Memory and Language,* 67, 4, 486-506. 10.1016/j.jml.2012.07.009

Perfors, A. (2016). Adult Regularization of Inconsistent Input Depends on Pragmatic Factors. *Language Learning and Development, 12:2*, 138-155, DOI: 10.1080/15475441.2015.1052449

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.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review, 21*(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

Plenio, M.B., & V. Vitelli (2001) The physics of forgetting: Landauer's erasure principle and information theory, *Contemporary Physics, 42*:1, 25-60, DOI: 10.1080/00107510010018916

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

Price, G. N., Bannerman, S. T., Viering, K., Narevicius, E., & Raizen, M. G. (2008). Single-photon atomic cooling. *Physical review letters, 100(9)*, 093004. https://doi.org/10.1103/PhysRevLett.100.093004

Prigogine I. (1978). Time, structure, and fluctuations. *Science* 201, 777–785. 10.1126/science.201.4358.777

Prigogine, I. & Stengers, I. (1984). *Order out of Chaos: Man's New Dialogue with Nature*. Flamingo Edition, London.

Protzner, A. B., Valiante, T. A., Kovacevic, N., McCormick, C., & McAndrews, M. P. (2010). Hippocampal signal complexity in mesial temporal lobe epilepsy: a noisy brain is a healthy brain. *Archives italiennes de biologie*, *148*(3), 289–297.

Radulescu, S. & Grama, I. (2021) Size Does Not Matter. Entropy Drives Rule Induction in Non-Adjacent Dependency Learning. *Manuscript under review.*

Radulescu, S., Giannopoulou, E., Avrutin, S., & Wijnen, F. (2021) Item-bound and Category-based Generalization. An Entropy Model. *Unpublished Manuscript.*

Radulescu, S., Kotsolakou, A., Wijnen, F., Avrutin, S. & Grama, I. (2021) Fast But Not Furious. When Sped Up Bit Rate of Information Drives Rule Induction. *Manuscript under review.*

Radulescu, S., Murali, M., Wijnen, F., & Avrutin, S. (2021) Turn That Noise On. Noisy Backgrounds Drive Rule Induction. *Unpublished Manuscript.*

Radulescu, S., Wijnen, F., & Avrutin, S. (2019). Patterns bit by bit. An entropy model for rule induction. *Language Learning and Development.* Advance online publication. https://doi.org/10.1080/15475441.2019.1695620

Radulescu, S., Wijnen, F., Avrutin, S., and Gervain, J. (2021) Same Processing Costs for Repetition and Non-Repetition Grammars in 6-month-olds: An fNIRS Study. *Unpublished Manuscript.*

Raizen M.G. (2009) Comprehensive control of atomic motion. *Science, 324,* 1403-06. DOI: 10.1126/science.1171506

Raven, J., Raven, J. C., & Court, J. H. (2000). Manual for Raven's progressive matrices and vocabulary scales. Section 3: The standard progressive matrices. Oxford, UK: Oxford Psychologists Press; San Antonio, TX: The Psychological Corporation.

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*(1), 30–54. https://doi.org/10.1016/j.cogpsych.2012.09.001

Reis, A.H. (2014). Use and validity of principles of extremum of entropy production in the study of complex systems. *Ann. Phys. 346*, 22–27

Reis, A.H. (2016). AD-HOC principles of "minimum energy expenditure" as corollaries of the constructal law. The cases of river basins and human vascular systems. *Int. J. Heat Technol. 34*, S147–S150

Richards, B.A., and Frankland, P.W. (2017). The Persistence and Transience of Memory. *Neuron, 94*, 1071-1084

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

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(6), 906–914.

Romberg, A. R., & Saffran, J. R. (2013). All together now: Concurrent learning of multiple structures in an artificial language. *Cognitive Science, 37*(7), 1290–1318. https://doi.org/10.1111/cogs.12050

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical Learning by 8-Month-Old Infants. *Science 274* (5294), 1926–1928.

Saldana, C., Smith, K., Kirby, S., & Culbertson, J. (2017). Is the strength of regularisation behaviour uniform across linguistic levels? In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1023-1028). Cognitive Science Society.

Samara, A., Smith, K., Brown, H., & Wonnacott, E. (2017). Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation. *Cognitive psychology*, *94*, 85–114. https://doi.org/10.1016/j.cogpsych.2017.02.004

Saults, J. S., & Cowan, N. (2007). A central capacity limit to the simultaneous storage of visual and auditory arrays in working memory. *Journal of Experimental Psychology: General, 136*(4), 663–684. https://doi.org/10.1037/0096-3445.136.4.663

Schneider, E. D., & Sagan, D. (2005). *Into the Cool: Energy Flow, Thermodynamics, and Life*. Chicago, IL: University of Chicago Press.

Schrödinger E. (1944). *What is Life?* Cambridge: Cambridge University Press.

Seidenberg, M. S., & Elman, J. L. (1999). Networks are not 'hidden rules'. *Trends in cognitive sciences*, *3*(8), 288–289. https://doi.org/10.1016/s1364-6613(99)01355-8

Sethna, J. (2006). *Statistical mechanics: entropy, order parameters, and complexity*, volume 14. Oxford University Press.

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

Sharma, V., & Annila, A. (2007). Natural process--natural selection. *Biophysical chemistry*, *127*(1-2), 123–128. https://doi.org/10.1016/j.bpc.2007.01.005

Sirois, S., Buckingham, D., & Shultz, T. R. (2000). Artificial grammar learning by infants: an auto-associator perspective. *Developmental Science, 3*(4), 442–456.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444–449. https://doi.org/10.1016/j.cognition.2010.06.004

Smith, S. M., & Handy, J. D. (2014). Effects of varied and constant environmental contexts on acquisition and retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(6), 1582–1593. https://doi.org/10.1037/xlm0000019

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

Szilard L. (1929). On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings. *Physik 53*, 124–133.

Thiessen, E. D., & Saffran, J. R. (2007). Learning to Learn: Infants' Acquisition of Stress-Based Strategies for Word Segmentation. *Language Learning and Development, 3*(1), 73–100. https://doi.org/10.1207/s15473341lld0301_3

Tincoff, R., & Jusczyk, P. W. (2012). Six-month-olds comprehend words that refer to parts of the body. *Infancy, 17*(4), 432–444. https://doi.org/10.1111/j.1532-7078.2011.00084.x

Tononi, G. (2008). Consciousness and Integrated Information: a Provisional Manifesto. *Biol. Bull. 215*, 216–242.
Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan & F. Restle (Eds.), *Cognitive theory, Vol. III* (pp. 200–239). Hillsdale: Erlbaum.

Townsend, J. T., & Eidels, A. (2011). Workload capacity spaces: a unified methodology for response time measures of efficiency as workload is varied. *Psychonomic bulletin & review*, *18*(4), 659–681. https://doi.org/10.3758/s13423-011-0106-9

Trambouze, P. (2006). Structuring Information and Entropy: Catalyst as Information Carrier. *Entropy*, *8*(3), 113–130. doi:10.3390/e8030113

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

Twomey, K. E., Ma, L., & Westermann, G. (2018). All the Right Noises: Background Variability Helps Early Word Learning. *Cognitive science*, *42 Suppl 2*(Suppl Suppl 2), 413–438. https://doi.org/10.1111/cogs.12539

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2009). Examining the relations among working memory capacity, attention control, and fluid intelligence from a dual-component framework. *Psychology Science, 51*(4), 388–402.

Van Egmond, M. (2018). Zipf's law in aphasic speech - An investigation of word frequency distributions. PhD dissertation, Utrecht University

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

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

Varpula S., Annila A., Beck C. (2013). Thoughts about thinking: cognition according to the second law of thermodynamics. *Adv. Stud. Biol. 5*, 135–149.

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019a). Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Processes, 56:3*, 229-255, DOI: 10.1080/0163853X.2018.1448677

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019b). Semantic Entropy in Language Comprehension. *Entropy*, *21*(12), 1159. doi:10.3390/e21121159

Verghese, P., & Pelli, D. G. (1992). The information capacity of visual attention. *Vision Research, 32*(5), 983–995. https://doi.org/10.1016/0042-6989(92)90040-P

Villarreal, D.M., Do, V., Haddad, E., and Derrick, B.E. (2002). NMDA receptor antagonists sustain LTP and spatial memory: active processes mediate LTP decay. *Nat. Neurosci.* 5, 48–52.

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.

Wagner, J. B., Fox, S. E., Tager-Flusberg, H., & Nelson, C. A. (2011). Neural processing of repetition and non-repetition grammars in 7- and 9-month-old infants. *Frontiers in psychology*, *2*, 168. https://doi.org/10.3389/fpsyg.2011.00168

Wang, F. H., Zevin, J., & Mintz, T. H. (2016). Learning Non-Adjacent Dependencies in Continuous Presentation of an Artificial Language. *Proceedings of the 38th Annual Conference of the Cognitive Science Society,* Austin, TX: Cognitive Science Society

Wang, F. H., Zevin, J., & Mintz, T. H. (2019). Successfully learning non-adjacent dependencies in a continuous artificial language stream. *Cognitive psychology*, *113*, 101223. https://doi.org/10.1016/j.cogpsych.2019.101223

Wilson, B., Spierings, M., Ravignani, A., Mueller, J. L., Mintz, T. H., Wijnen, F., van der Kant, A., Smith, K., & Rey, A. (2020). Non-adjacent Dependency Learning in Humans and Other Animals. *Topics in cognitive science*, *12*(3), 843–858. https://doi.org/10.1111/tops.12381

Wonnacott, E. (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language* 65, 1–14.

Wonnacott, E., & Newport, E.L. (2005). Novelty and regularization: The effect of novel instances on rule formation. In A. Brugos, M.R. Clark-Cotton, and S. Ha (eds.), *BUCLD 29: Proceedings of the 29th Annual Boston University Conference on Language Development.* Somerville, MA: Cascadilla Press.

Wu, T., Dufford, A. J., Mackie, M. A., Egan, L. J., & Fan, J. (2016). The Capacity of Cognitive Control Estimated from a Perceptual Decision Making Task. *Scientific reports*, *6*, 34025. https://doi.org/10.1038/srep34025

Yufik Y. M. (2013). Understanding, consciousness and thermodynamics of cognition. *Chaos Solitons Fractals* 55, 44–59. 10.1016/j.chaos.2013.04.010

Zipf, G.K. (1949). *Human behavior and the principle of least effort*, Addison-Wesley Press

Samenvatting (Summary in Dutch)

Dit proefschrift is een verzameling artikelen die de resultaten presenteren van een onderzoeksproject waarin het leren van taalregels vanuit een informatietheoretisch perspectief is onderzocht. Het belangrijkste doel van dit onderzoeksproject was om een innovatief entropiemodel voor regelinductie voor te stellen en te testen, gebaseerd op Shannon's 'noisy-channel' (ruiskanaal) coderingstheorie (Shannon, 1948).

Regelinductie (generalisatie of regularisatie van regels) is een essentieel mechanisme voor taalverwerving dat taalleerders in staat stelt niet alleen specifieke items (bijv. fonemen, woorden) te onthouden wanneer ze worden blootgesteld aan linguïstische input (taal), maar ook om relaties tussen deze items te leren. Deze relaties variëren van statistische patronen tussen specifieke items die aanwezig zijn in de linguïstische input (Saffran, Aslin, & Newport, 1996; Thiessen & Saffran, 2007) tot meer abstracte categorie-/regelinductie (Marcus, Vijayan, Rao, & Vishton, 1999; Smith & Wonnacott, 2010; Wonnacott, 2011; Wonnacott & Newport, 2005). Taalleerders onthouden bijvoorbeeld niet alleen woorden en combinaties van woorden, zoals *moeder liep langzaam* en *vader praatte vriendelijk*, maar ze leiden ook generalisaties (regels) af zoals 'lach *-te* of lach *-end*' aan specifieke items om uit te drukken een handeling uit het verleden of de manier waarop een handeling wordt uitgevoerd. Bovendien zijn leerders in staat af te wijken van specifieke combinaties van items en vormen ze zo categorieën en algemene regels: bijvoorbeeld, Zelfstandig Naamwoord-Werkwoord-Bijwoord is een goed gevormde reeks, waarbij elke categorie een vrijwel oneindig aantal specifieke items kan bevatten. Dit onderzoeksproject richtte zich op de inductieve stappen van het onthouden van specifieke items, naar het afleiden van regels (of statistische patronen) tussen deze specifieke items (in onze terminologie: *itemgebonden generalisaties*), en ook naar het vormen van regels die van toepassing zijn op categorieën van items (*categoriegebaseerde generalisatie*).

De belangrijkste onderzoeksvragen van dit onderzoeksproject zijn de volgende:

1. Zijn de twee vormen van generalisatie de uitkomst van twee afzonderlijke mechanismen, waarbij *statistisch leren* resulteert in *itemgebonden generalisaties* op een lager niveau, en het *abstract leren van regels* leidt tot de op *categorie gebaseerde generalisaties* van hogere orde? Of zijn het uitkomsten van hetzelfde mechanisme? Als het maar één mechanisme is, is het dan een gefaseerd mechanisme dat geleidelijk overgaat van de ene vorm van generalisatie naar de andere? Of is het een abrupte verschuiving?

2. Wat zijn de factoren die de verandering van *itemgebonden* naar *categoriegebaseerde generalisatie* sturen?

Om deze vragen te beantwoorden, presenteert dit proefschrift een nieuw entropie- en ruiskanaalcapaciteitsmodel (kortom, entropiemodel) voor regelinductie, dat is gebaseerd op Shannons *ruiskanaalcoderingstheorie* (Shannon, 1948). *Entropie* is een informatietheoretische maatstaf voor de hoeveelheid en complexiteit van informatie, terwijl *kanaalcapaciteit* de hoeveelheid informatie (inclusief ruis) is die per tijdseenheid kan worden overgebracht.

De belangrijkste hypothese van het entropiemodel is dat regelinductie een coderingsmechanisme is dat *geleidelijk* wordt aangedreven door de dynamiek tussen een externe factor – *invoer-entropie* – en een interne factor – *kanaalcapaciteit*. We definiëren de coderingscapaciteit van onze hersenen als *kanaalcapaciteit* op computationeel niveau, in de zin van Marr (1982), wat de eindige snelheid is van informatiecodering (bits per seconde). Op algoritmisch niveau kan de *kanaalcapaciteit* worden ondersteund door cognitieve capaciteiten die betrokken zijn bij het verwerken en coderen van informatie, zoals geheugen en aandacht.

Hoofdstuk 1 onderzocht het effect van de eerste factor van het entropiemodel – *inputentropie* – op regelinductie. In twee artificiële grammatica-experimenten werden volwassenen blootgesteld aan een op herhaling gebaseerde XXY-grammatica van drie lettergrepen (bijv. *da:-da:-li*), in zes experimentele condities met toenemende *input-entropie*. In het geval van een XXY-grammatica betekent *itemgebonden generalisatie* het afleiden van een *gelijk-gelijk-verschillende* regel alleen met bekende lettergrepen uit de aangeboden stimuli. Er werd verondersteld dat een toename in entropie de tendens naar een meer abstracte, op categorieën gebaseerde generalisatie versterkt (d.w.z. een *gelijk-gelijk-verschillende* regel met ook onbekende lettergrepen). De resultaten toonden aan dat wanneer *inputentropie* toeneemt, de neiging om naar *categoriegebaseerde generalisatie* te gaan *geleidelijk* toeneemt, wat bewijs levert in het voordeel van ons entropiemodel.

In Hoofdstuk 2 hebben we een onderzoeksvraag behandeld in de context van vroege taalontwikkeling zoals voorspeld door het entropiemodel. Omdat de *kanaalcapaciteit* wordt ondersteund door cognitieve capaciteiten, zoals geheugen, die zich ontwikkelen met leeftijd, wordt aangenomen dat baby's een verminderde *kanaalcapaciteit* hebben in vergelijking met volwassenen. Zo wordt voorspeld dat de neiging van baby's tot regelinductie wordt gedreven door minder *inputentropie* dan de volwassenen.

We hebben met behulp van functionele 'near-infrared spectroscopy' (fNIRS) getest of, én hoe zes maanden oude baby's op herhaling gebaseerde, taalkundige regelmatigheden (ABB, bijv. "bu ra ra") verwerken in vergelijking met niet-herhalende sequenties (ABC, bijv. "bu fa zo"), waarbij we de invoerentropie (laag versus hoog) manipuleerden. Er werd voorspeld dat baby's in staat zouden zijn om zowel ABB- als ABC-sequenties te verwerken, en ook om onderscheid te maken tussen deze sequenties, maar dat ze dit gemakkelijker zouden doen onder omstandigheden met hoge entropie.

We vonden een trend naar hogere hersenactivatie voor niet-herhalende sequenties, en ook hogere activatie bij hoge entropie. We vonden echter geen verschil tussen de twee grammatica's, noch vonden we een verschil tussen de lage en hoge entropiecondities. Deze resultaten suggereren dat baby's van zes maanden in staat zijn om zowel de herhalende als de niet-herhalende patronen te verwerken, waarbij de verwerkingskosten hetzelfde zijn voor beide patronen. Onze bevindingen zijn de eerste die een ontwikkelingsverandering in taalverwerving onthullen tussen de leeftijd van zes maanden en de geboorte, waarvoor eerder discriminatie tussen herhalende en niet-herhalende patronen werd gevonden (Gervain et al., 2008).

Hoofdstuk 3 onderzocht de geleidelijke overgang van 'uit het hoofd leren' naar *itemgebonden generalisatie* en *categoriegebaseerde generalisatie*, zoals verondersteld door het entropiemodel. Om dit te onderzoeken, hebben we volwassenen blootgesteld aan een lage en een gemiddelde entropieversie van de XXY-grammatica (uit hoofdstuk 1), en hebben we de hypothese getest dat lage inputentropie niet alleen het uit het hoofd leren van specifieke items en hun waarschijnlijkheidsverdeling, zoals aanwezig in de input, vergemakkelijkt, maar ook *itemgebonden generalisatie*. Ook hebben we gekeken naar individuele verschillen in specifieke componenten van de cognitieve capaciteiten waarvan we veronderstelden dat ze ten grondslag liggen aan *kanaalcapaciteit*, oftewel expliciete/impliciete geheugencapaciteit en domeinbrede patroonherkenningsvaardigheid, die op de werkgeheugencapaciteit berust.

Onze bevindingen tonen aan dat een lage *inputentropie item-gebonden generalisatie* inderdaad vergemakkelijkt en niet enkel het leren van specifieke items. We vonden ook dat een toename van *de input-entropie* leidt tot een meer *categoriegebaseerde generalisatie*. Bovendien vonden we dat in de conditie met gemiddelde entropie, maar niet in de conditie met lage entropie, leerders met een lage incidentele geheugencapaciteit, maar met een hoge domeinbrede patroonherkenningsvaardigheid een grotere neiging vertonen tot *categoriegebaseerde generalisatie* dan leerders met een hoge incidentele memorisatiecapaciteit, maar met een lage domeinbrede patroonherkenningsvaardigheid. Deze bevindingen ondersteunen ons entropiemodel.

In Hoofdstuk 4 hebben we ons entropiemodel voor regelinductie verder uitgebreid van een op herhaling gebaseerde XXY-grammatica naar een complexere a_iXb_i-grammatica met niet-aangrenzende (gerelateerde) elementen (bijv. *rak_naspu_tuf*). In dit type grammatica voorspelt een specifiek item *a* (bijv. *rak*) altijd een specifiek item *b* (bijv. *tuf*), en vormen een vast $a_i\_b_i$-frame met een tussenliggend element uit de categorie van een grote variatie aan X'en. We veronderstelden dat, hoewel hoge *input-entropie categoriegebaseerde generalisatie* voor de X-categorie bevordert, het de *itemgebonden generalisatie* voor de specifieke $a_i\_b_i$-afhankelijkheden hindert. Het effect van een toename in de entropie op het leren van dit type grammatica is dus niet een *geleidelijk* betere prestatie zoals we vonden voor de XXY-grammatica (Hoofdstuk 1 en 3). Ons entropiemodel verfijnt eerdere theorieën die beweerden dat een grote setgrootte van de tussenliggende X'en tot het beter leren van dit soort niet-aangrenzende afhankelijkheden zou leiden (Gómez, 2002; Gómez & Maye, 2005). We veronderstelden dat niet alleen de setgrootte is, maar de *inputentropie* het leren moduleert.

We hebben volwassenen blootgesteld aan drie entropiecondities – laag, gemiddeld en hoog – van een niet-aangrenzende afhankelijkheid a_iXb_i-grammatica vergelijkbaar met die van Gómez (2002), terwijl de setgrootte van tussenliggende X'en gelijk werd gehouden. Zoals voorspeld, zagen we dat deelnemers niet-aangrenzende afhankelijkheden beter leerden en generaliseerden in omstandigheden met de hoogste entropie dan in omstandigheden met gemiddelde en lage entropie. Bovendien vonden we een U-vormig patroon in het leren van niet-aangrenzende afhankelijkheden als functie van toenemende *inputentropie*, zonder dat we bewijs zagen voor het succesvol leren in de medium-entropieconditie, wat in lijn is met eerdere bevindingen (Onnis et al., 2003; 2004).

In Hoofdstuk 5 hebben we eerst theoretisch uiteengezet hoe *kanaalcapaciteit* en snelheid van informatieoverdracht kunnen worden geschat in een kunstmatige taalleeromgeving voor regelinductie. Daarna hebben we de tijdsvariabele van de *kanaalcapaciteit* direct gemanipuleerd in twee andere artificiële grammatica-experimenten met volwassenen.

In het bijzonder hebben we de bitsnelheid van informatieoverdracht verhoogd niet door simpelweg de tijd tussen stimuli met een willekeurige hoeveelheid te verminderen, maar met een factor die we hebben berekend op basis van gegevens uit onze eerdere experimenten (hoofdstuk 1), door gebruik te maken van de *kanaalcapaciteit* formule. In het eerste experiment stelden we volwassenen bloot aan de laagste entropieversie van de XXY-grammatica uit hoofdstuk 1, ofwel in een conditie met lage transmissiesnelheid of een met hoge transmissiesnelheid. In het tweede experiment stelden we volwassenen bloot aan een lage entropieconditie van de a_iXb_i-grammatica (uit hoofdstuk 4), de ene groep met een lage transmissiesnelheid en een andere groep met hoge transmissiesnelheid.

We vonden dat wanneer we de bitsnelheid van informatieoverdracht in een op herhaling gebaseerde XXY-grammatica verhoogden, de neiging van leerders tot *categoriegebaseerde generalisatie* toenam, zoals voorspeld door ons model. Omgekeerd vonden we dat een verhoogde bitsnelheid van informatieoverdracht in een complexere *aXb*-grammatica met niet-aangrenzende afhankelijkheden in het algemeen leidde tot slechtere leerresultaten. Deze uitkomsten zijn in lijn met de voorspellingen van ons entropiemodel.

In Hoofdstuk 6 hebben we gekeken naar het effect van de ruis-kanaalcapaciteit (Shannon, 1948), door *ruis* (d.w.z. willekeurig stimulus-irrelevant materiaal) op de achtergrond toe te voegen, terwijl we de deelnemers blootstelden aan de laagste entropieversie van de XXY-grammatica uit Hoofdstuk 1. In de ene conditie hadden de leerders een extra taak naast het luisteren naar de XXY-taal, namelijk het letten op en onthouden van specifieke items uit het geluidsmateriaal (simultaantaakconditie), terwijl deelnemers in een andere conditie geen extra taak kregen tijdens de ruis (afleidingsconditie).

We ontdekten dat toegevoegde signaal-irrelevante entropie (ruis) de neiging naar *categoriegebaseerde generalisatie* bevordert, ongeacht de lage entropie van het signaal in de input, maar cruciaal enkel wanneer er geen extra taak volbracht hoefde te worden tijdens het ruismateriaal.

Hoofdstuk 7 schetst het eerste gezamenlijke informatietheoretische en thermodynamische model van regelinductie. Concreet suggereerden we met dit innovatieve perspectief dat de tweede wet van de thermodynamica de vraag kan beantwoorden waarom regelinductie plaatsvindt, terwijl de constructieve wet van de thermodynamica de vraag kan beantwoorden hoe regelinductie plaatsvindt.

About the author

Silvia Radulescu was born in Ploiești, Romania. Between 1999 and 2003, she studied Philology at the University of Letters in Bucharest and graduated with a Major in Romanian Language and Literature, and a Minor in English Language and Literature. In 2011, she started a 2-year Research Master program at Utrecht University, entitled *Linguistics – the Study of the Language Faculty.* She graduated with a Major in Psycholinguistics and obtained a Master's Degree *cum laude* in Linguistics (Research). After graduating, she started her own scientific research on the entropy model presented in this dissertation as a guest PhD candidate at the Utrecht University. In 2016, she was awarded a 4-year research grant by the Netherlands Organization for Scientific Research (NWO) to continue her research as a PhD candidate at Utrecht University. The results of this research program are presented in this dissertation.