Between logic and language: a quest for primitives of thought

Investigations on the origin of logical operators and their relationship with language

Irene Canudas Grabolosa

TESI DOCTORAL UPF / 2021

Thesis Supervisors

Prof. Luca Lorenzo Bonatti, Dr. Elena Pagliarini

Department of Information and Communication Technologies

Universitat
Pompeu Fabra
Barcelona

A la meva família

Acknowledgements

A thesis is never, ever! something you do on your own. This is why I need to cite many people, and I apologize beforehand to all of those I will forget to mention.

First of all, I would like to acknowledge my supervisor, Luca. His guidance and supervision in every step of the process have been essential. He has been revising everything I did, maybe taking his time, but always with enormous respect and a pinch of humor, for which I am very grateful. I'd also like to say a big thanks to Elena Pagliarini, who's been my supervisor in the shadows first and in the light after, and who's spent tons of time and efforts correcting, suggesting, and finding participants. Also, Gennaro Chierchia deeply contributed to this work; he kindly explained his theory and discussed our results with us, helping us make sense of them.

Before moving on, I would also like to thank the members of my tribunal Dr. Justin Halberda, Dr. Teresa Guasti, Dr. David Barner, Dr. Salvador Soto-Farraco and Dr. Kalinka Timmer, for taking the time to read and evaluate this work.

I would also like to thank all people who improved this thesis without knowing me personally, by stopping by at my posters and commenting on them -I can not name all of them, but their comments were wise and greatly helpful. I need to especially thank JR. Hochmann and N. Cesana-Arlotti, who besides commenting on a scientific level, helped me in my desperate search of online preschooler participants. Cristina, Núria, Kinga, Ana, and Aina, from the RICO group, have provided me useful discussions, suggestions, and tips. També vull agrair especialment l'ajuda d'en Xavi i la Sílvia, que m'han ajudat tècnicament i a l'hora de buscar participants (gràcies a l'Erik, també, que per culpa de la seva mare li va tocar ser pilot). Gràcies a la Cristina i la Kàtia. I la Cinta, que sense el seu cop de mà no hauria acabat mai, mai, mai l'estudi en línia. I a en Simone, l'autor de tots els vídeos de Maya!

I must have a warm word for Daniele Panizza and Anna Martinez-Alvarez, who besides their crazy love for bird songs, gave me also a hand with participants and ideas, as well as providing scientific suggestions, political discussions, and the latest gossips in the linguistic field.

I am also very grateful to the old good extended Computational Neuroscience Group, who 'adopted' me in their lunches, coffee breaks, and barbecues even though I have no degree in Physics: Andrea (the coffee master), Mattieu, Gorka, Ane, Ludovica, Laura, Kat, Xenia, Victor. Also, Sofia, who was adopted as well.

And Sander and Irene, not part of the CNG but are real artists. And Sophie, Veish and old Venera, who were there just a semester but they made it unforgettable.

Kalinka and Jess were also decisive in this process. Kalinka appeared in the last half but she's grown to be unreplaceable in all aspects of life. Y Jessica, que es mi hermana de doctorado, compañera de promoción y que no creo que imagine hasta qué punto esa tesis ha sido posible gracias a ella, sus conversaciones, pausas de café y neuras comunes.

Haig també de donar les gràcies a tots els meus amics que han patit aquest procés, amb més o menys intensitat. Gent musical de la Fossa Pupília: Laura, Bet, Cèlia, Ferran, Jan, Mar, Marta M, Marta F, Uri, Clara; gràcies per totes les excursions, viatges, sopars, pel·lícules, converses i confabulacions (també polítiques) plegats. L'Anna i la Cati també m'han seguit de prop durant tots aquests anys, i els ha tocat el rebre tot sovint; però no sé per què encara segueixen volent-me veure! L'Ada i la Clàudia, LES companyes de pis en majúscules, i també la Meri, mereixen una menció d'honor en aquesta llista; així com la Laia i la Núria, amb qui pràcticament hem crescut plegades i encara tenen paciència amb mi.

Devo ringraziare anche Rosa, Maria Assunta, Roberto, Natale, Davide (sì, Da, parlo ancora delle stesse grafiche), per tutto l'aiuto, gentilezza e comprensione ogni volta che sono stata in Italia e non mi sono praticamente alzata del tavolo della sala. Anche Rosanna (ed Elena) devo ringraziarle specialmente per aversi prestato a essere volontarie in uno studio oggettivamente noioso. E Junior per le registrazioni professionali che ha fatto per me (ti devo ancora una pizza e non me ne sono dimenticata).

Evidentemente, devo ringraziare anche Carmine, che è sempre vicino (vabbè, non fisicamente, si sa) e mi ha sempre sostenuta. Diciamo solo che la parte meno importante di tutto quello che ha fatto è stata la revisione di quei almeno 10.000 testi in italiano...

I finalment, a tota la meva família; els meus pares, germans, cosins, tiets i avis: Rosa i Eduard, Pau i Judit, Núria, Adrià, Jana, Aniol i Mariona, Assumpta, Martí, Montse, Bassi, Sònia, Jordi, Maria, Fina. No crec que tingui prou paraules per agrair-vos tot el que heu fet per a mi, des d'alimentar-me a fer-me riure i sempre, sempre estar al meu costat (fins i tot si calia fer *vodú*, oi, Pau?). Un agraïment especial als meus pares i germans, a la Mariona, que com l'Elena va ser víctima d'un estudi més aviat avorrit, i a l'Adrià, que em va gravar desinteressadament cada vegada que li vaig demanar, la Montse que m'ha corregit totes les cites i la bibliografia de la tesi (i mira que hi havia feina!), i a la Rosa, la Sònia, en Jordi i l'Eduard que van distribuir els meus estudis insistentment.

I evidentment a totes les persones que han participat en els meus estudis, inclosa l'escola Escorial (Vic), i les vora de 150 famílies que van participar amb els seus fills en algun dels estudis.

A tots vosaltres, moltes, moltes gràcies.

Abstract

In this thesis we focus on the relationship between language and logic, to try to find potential primitives of thought.

We depart from a well-established linguistic theory (the theory of exhaustification), which claims a higher importance of logic in the understanding of language, and we first focus on a paradigmatic case included in the theory, the meaning of words with logical valence. We studied their development in preschoolers and the systematicity of their meanings according to the logical context in adults. In the second part of the experimental work, we investigate a vital subcomponent of the theory: how entailment relationships are understood by preschoolers and adults. Finally, we take the first steps to study entailment relationships non-linguistically.

Our results indicate a pervasive role of logic in language and open up a promising line of research in which language is used to access thought.

Resum

En aquesta tesi ens centrem en la relació entre el llenguatge i la lògica, per intentar trobar capacitats de raonament inherents a l'ésser humà.

Partim d'una teoria lingüística ben establerta (la teoria l'exhaustifiació), que reivindica un pes superior de la lògica en la comprensió del llenguatge. Primer ens centrem en un cas paradigmàtic inclòs en l'esmentada teoria, la comprensió de paraules amb equivalència lògica, fixant-nos en el seu desenvolupament en preescolars i la sistematicitat dels seus significats segons el context lògic en adults. Seguidament, investiguem un subcomponent vital de la teoria: com entenen les relacions de conseqüència lògica els infants d'edat preescolar i els adults. Finalment, fem els primers passos per estudiar aquestes relacions sense recórrer a la llengua.

Els nostres resultats indiquen un paper generalitzat de la lògica en el llenguatge i obren una línia de recerca prometedora en què s'usa el llenguatge per accedir al pensament.

Preface

There is a town, one hour and a half ride by car from Barcelona called *Espluga de Francolí*. It is a peaceful, bucolic town, with its church, modest houses spread around it, vividly contrasting with the vineyards that grow in its surroundings. There is nothing, from first sight, that would tear it apart from other quiet villages nearby and around the Mediterranean basin. But yet, it has something that makes it very special.

It has a cave.

Let's imagine you can enter this cave -even if now it is forbidden. If you could, you would enter another world. You would go back 15,000 years in time and discover, carefully carved on the walls and roof of the cave, hundreds of dozens of mares, does, bows, and rabbits, in what is one of the most outstanding prehistorical sites of the Mediterranean region.

While you observe the paintings, some of them quite abstract, some very realistic, you would probably ask yourself 'Why did they do that?. Why did they feel the need to *decorate* the walls in such a way?' And perhaps, if you think a little more you will face another interesting question: why were they *even able* to paint? Where did their ability come from?

We can admire these paintings and carvings only by their beauty, but in fact, they are quiet witnesses of something more intricate: the fact that our ancestors were able to plan on extremely abstract levels and execute actions whose goal was purely symbolic, far beyond the basic needs of the individuals.

Actually, we do not need to talk only about paintings. We can think about the impressive amounts of technology and inventions the human species has created to adapt the environment to its needs. We can see some of these tools in our cave, but virtually anything we interact with has been thought and designed by another human being -sometimes by a machine invented by another human being.

That trait differentiates us from the rest of the animals in the animal kingdom: we can dramatically transform our environment and adapt it to our needs. There have been countless studies showing that different non-human animals can perform impressive complex tasks, but we are yet to find a kangaroo that invents a rocket.

What are these precise abilities that make our minds so powerful? Where do they come from? These questions have fascinated philosophers since the dawn of time;

and have attracted the interest of developmental and comparative psychologists, linguists, engineers, and artificial intelligence professionals.

All of them were puzzled by human's unique conceptual world, able to conceptualize and work with general, abstract concepts in a highly productive way.

At the same time, the human species has yet another characteristic that differentiates it from the other animals: language. Humans transmit thoughts, products of our conceptual world, to other individuals using language. In doing that, we create a common mental stage via language, and we give room for shared thoughts and, thus, shared knowledge.

Language, at the basis of this communication process, is, in fact, a bizarre product. Studied under multiple focuses, it presents itself as a powerful combinatorial mechanism, which some place at the very beginning of thought itself, while others consider a byproduct. The reality is that it seems to be a very valid entrance door to a highly hidden domain, that of concepts.

In this thesis, we will try to use this door to access the thought. We will consider language and ask ourselves how to trace certain of its characteristics back to the thought, in the umpteenth attempt to uncover the complex relationship between these two extraordinary human characteristics. The result is, needless to say, not the answer to any of our deep philosophical questions; rather just a new humble attempt to contribute to unveil what makes humans *human*.

Manlleu, Autumn 2020

Contents

List of Figures

Chapter 0

Introduction

The human mind is unique in the animal world. It has a large capacity to represent objects, events or situations; to analyze abstract regularities, interpretations of others' goals and thoughts…; and draw conclusions and predictions based on these representations, in a short time.

Thought bears upon another deeply human capacity: language. However, while the relationship between language and thought seems undeniable, the amount of conceptual resources they share is more debated. All across the history, philosophers, psychologists, linguists and now cognitive scientists have argued for the opposite horns of this relation. The spectrum has ranged from those who conceive language as the source of concepts (Quinean linguistic determinism, Van Orman, 1960) to those who deny any effect of language on thought (Frege, 1918; Fodor, 1989). The first school of thought tends to understand adult representations of the world as a cultural construct (learned through natural language), and, as a such, a construct that can change from generation to generation and from culture to culture (hence the Sapir-Worf hypothesis, Boroditsky, 2006). As for the second school, not only does it deny linguistic determinism but it also understands language as derivative from thought, rather than the other way around (Fodor, 1989).

The main unsolved question, therefore, can be summarized with this simple yet complex question: 'Does language change the way humans think?', or, even more broadly, 'What has been the role of language in human evolution?'. Unfortunately, such questions cannot be directly addressed by archaeology or evolutionary biology: we simply have no way to access the relevant information. Nor did extensive studies aimed at training nonhuman animals to learn language help (Seidenberg &

Petitto, 1979). And of course they cannot be addressed by studying adult humans, because they already have language. The most sensible way to address that big question has been studying human infants: humans are the only known species with full capacities to develop language, and they develop most of it the few years from infancy to 3-4 years of age. The question therefore, can undergo a subtle change: ‘does language acquisition lead infants to think differently?’.

The topic of reasoning capacities in infancy has been traditionally overlooked because of the influence of one of the most prominent psychological theories of human development. Its author, Piaget, argued that children could not fully reason logically before adolescence (Piaget & Inhelder, 1951). However, in recent years, several studies have challenged this assumption (Feigenson & Halberda, 2004; Cesana-Arlotti et al., 2018), showing that some form of logical thinking may exist before humans even start to talk or walk. In parallel to this recent line of research on logical thinking in infancy, a growing number of studies has been investigating the logicality of language, and how logical structures that confer the meaning of linguistic expressions are treated and acquired by children (Beilin & Lust, 1975b; Braine & Rumain, 1981; Chierchia, Crain, Guasti, Gualmini, & Meroni, 2001; Noveck, 2001; Papafragou & Musolino, 2003; Guasti et al., 2005; Hurewitz, Papafragou, Gleitman, & Gelman, 2006; Barner, Brooks, & Bale, 2011; Foppolo, Guasti, & Chierchia, 2012; Gualmini, 2014; Tieu, Romoli, Zhou, & Crain, 2016; Pagliarini, Bill, Romoli, Tieu, & Crain, 2018). However, both lines have mainly ignored developments in the other. Therefore, the exact relationship (and potential causal directionality) between logic and language still remains veiled.

In this research we try to overcome the lack of exchange between different areas relevant to the understanding of this relationship by focusing on the logicality of language as a tool to obtain information about basic primitives of thought. Our long term aim is to identify components for several logical computations in language and investigate their psychological reality in development.

Specifically, in this work we will focus on words with logical valence, their acquisition and their meaning. Those are words such as ‘some’, ‘all’, and connectives like ‘and’ or ‘or’, which are thought to be closely tied to their meaning in standard logic, $\exists$, $\forall$, $\wedge$, and $\vee$, respectively. There is a longstanding open issue on how the meanings of these words are acquired. They are famous for having an *at least* meaning (a ‘literal’ meaning, the one given by classical logic) but also an enriched meaning, derived from the *at least* one (see example (2), below):

(1) If you open the door, I’ll take some animals outside.

a. ***at least* meaning**: I'll take more than one animal (possibly all) outside.

b. **Enriched meaning**: I'll take more than one animal, but not all, outside.

(2) If you open the door, I'll take the cat or the dog outside.

a. ***at least* meaning**: I'll take the cat or I'll take the dog or I'll take both animals outside.

b. **Enriched meaning**: I'll take the cat or I'll take the dog but I won't take both animals outside.

However, the reason why these different meanings are available and their relative importance in acquisition are poorly understood. Classically, this phenomenon is described as an inferential process following Gricean conversational norms (Grice, 1975; Horn, 1989): given a sentence, you apply pragmatic norms to restrict its meaning. This accords to the traditionally attitude to describe the comprehension of logical words as stemming from the interaction of semantics (part of the 'hard-core' of language) and pragmatics (where non-specific mechanisms apply).

There is no denying of the importance of pragmatics in understanding. However, recent studies have proposed to analyze the phenomenon we described above entirely as gramatic, without the need to appeal to pragmatic inferences (Fox, 2007; Chierchia, 2013). These proposals are based on the idea that each sentence is understood as contrasted against a set of relevant alternative sentences, given both by the context and the lexical entry of the world. By a process known as *exhaustification*, all other possible meanings (alternative sentences) are negated, and thus the real meaning is achieved.

Some of the computations needed for exhaustification resemble abilities infants have been shown to have in unrelated contexts. This is so for example, for their ability to hierarchically organize sets (Feigenson & Halberda, 2004), or their ability to acquire novel information by exploring alternative hypotheses (Stahl & Feigenson, 2015). However, our understanding of the required computations is partial and, for some of these computations, inexistent, even though they constitute a useful case-study to find candidates to primitives of thought.

0.1 Goals of the current research

The present research has two grand aims. First, it seeks to investigate the logical capacities supporting the basic understanding of a crucial part of language: that involving logic. As we have anticipated, such an enterprise departs from a well-established theory about the logicality of language and exploits its subcomponents, what we have judged to be the basic logical processes behind it, to obtain information about basic primitives of thought.

Indeed this is our second, long-term goal: to describe potential candidates to primitives of thought. We stress that we are not interested in the theory of exhaustification *per se*. Instead, we use it as *a tool* in an attempt to get at the deepest part of the logical components from which a linguistic system is based or can emerge: we are interested in finding, departing from language, which logical computations are available to children and how they work. In plain language, we are interested in the necessary logical structure of *thoughts* rather than in the necessary logical structure of *language*.

More specifically in this thesis, we will focus on two essential components of the linguistic theory that have also a prominent role in the inferential reasoning abilities: finding intersections between the development and meaning of words with logical valence, and investigating the role of entailment in children's abilities.

Studying these two components, and specifically clarifying their intersections could explain both the linguistic phenomena related to exhaustification and the inferential reasoning abilities which are so crucial for our species to develop the extended network of knowledge and theories that uniquely characterizes us.

This thesis is organized as follows: The first chapter (chapter 1) is a revision of the critical theories and evidence that serve as foundation for our work.

The following two chapters (chapters 2 and 3) will be devoted to the analysis of the semantic and psychological role of words with logical valence. A better understanding of these phenomena will help to appreciate the relation between language and thought. We will investigate the meeting points of connectives and numerals, focusing on how they are affected by exhaustification, and the abilities to conceive and handle restricted domains of reference.

In chapter 2, we will begin by examining how equivalent words with logical valence (numerals and their equivalent conjunctions) are acquired and what children's understanding of contexts involving them can tell us something about a possible shared origin. Chapter 3 will move a step further to study how these logical operators are expressed in language, exploring their differences and similitudes when they involve the same operation of exhaustification.

Chapters 4 and 5 will deal with the understanding of logical entailment, as a step for seriation abilities, needed to order the alternatives in the theory. Indeed, entailment is an important logical operation per se, and its role in the understanding of language is not debated (Horn, 1989; Gualmini, 2014). Therefore, its existence can be considered as a case to test whether thought exists before or after language arises: if preverbal infants do not seem sensitive to the basic operations necessary to implement it, this would be a strong argument that language plays a causal role for their genesis. If instead, these are present before infants learn to speak, then one could conclude that the roots of these processes can be traced before language emerges, thus contributing with a shred of non-conclusive but interesting evidence favoring the theory that logical thought is independent of language.

Chapter 4 will take up this endeavor by cross-linguistically investigating children's understanding of entailment relations and inferences in a linguistic context. Finally, in chapter 5, we will present a first attempt to reveal non-linguistic entailment abilities in adults as a preliminary step to uncover them in preverbal infants.

This thesis is an attempt to move from a precise theoretical space to an empirical one by operationalizing the theoretical concepts. We will try to bring together two lines of parallel research, linguistics and developmental cognitive science, wishing that the linking of the two will benefit both: providing a much stronger theoretical framework for cognitive science while filling the semantic theory with evidence about the roots of its theoretical proposals.

Our grand plan had to be adjusted by the turmoils that hit the world in this past year, and particularly hit behavioral research with children and infants. We had to scale back some of the planned avenues of research that we deemed necessary to realize our plan. Still, we hope that, with all its limits, our research will contribute to the understanding of the ontogenesis of the relation between human reasoning and language.

Chapter 1

Theoretical background

The present chapter will present the theoretical background necessary to understand the experimental work we expose in the following chapters.

The topic we are dealing with has been around for a long time. Ever since Plato, philosophers have tried to understand how to decipher thought, elucubrating on the relation between it and language. A radical shift occurred when formal logic was invented: the new formal tools suggested to some to entirely dissociate thought from natural language and uniquely make use of formal languages and logic in describing thoughts, since natural language was considered to be too confusing and ambiguous. We briefly present this line in subchapter 1.1.

This characterization changed with Montague. He showed that natural language could be described in mathematical terms, exactly as formal languages were. Montague's ideas, along with cognitive science's aim to describe language in the mind, opened up the way to consider natural language as a valid entrance gate to thought (subchapter 1.2.1).

Within this view of language, the border within pragmatics (that is, the use of language) and semantics (the meaning) acquired a renewed importance: pragmatics had traditionally assumed all idiosyncratic and cross-cultural differences, while semantics was thought to be universal. Where to trace the line was, therefore, of importance. A paradigmatic case study in the grey space between the two rapidly became that of words with direct logical correspondence. These words (quantifiers, connectives, numerals) show a duplicity of meanings in natural language use: they have an *at least* meaning roughly corresponding to the meaning they acquire in logical and formal languages (which would correspond to the 'basic meaning'), along with an enriched, more restricted meaning, which is supposed to come from pragmatic reasoning: semantics and pragmatics interacting in the same words (see subchapter 1.2.2).

Recently, some linguistic theories have suggested that the enriched meaning of these words also comes from semantics and not from pragmatics, via computations that can be made formally rigorous. Interestingly, these computations have great explanatory power, since they also extend to other linguistic phenomena (such as the syntactic distribution of 'any') thought to be independent of pragmatics. We will present this theory in subchapter 1.2.3. If the theory is true, there would be more structure and compositionality in natural language that what had been previously assumed. And, importantly for our research, this structure would potentially be closer to, and perhaps have direct resonance to, the structure of thought, our target in this inquiry.

In order to investigate the psychological validity of the theory, in section 1.2.4 we identify four main subcomponents of the proposed computations: retrieval of the logical meaning of the word, generation of a set of alternatives, ordering of these alternatives and closure of the set.

In the last section of this chapter, section 1.3, we review what is known of each of the four subcomponents in thought, outside the realm of linguistics. An effective way, one of the few available, to investigate thought without linguistic interferences is to turn to infants' cognitive capacities, since infants do not speak yet. That is why we center the last section of the chapter to investigate what is known about infants' abilities and learning related to our four subcomponents.

We start by an overview of the similitudes that inferential processes available to infants bear to the computations we are discussing (section 1.3.1).

We then turn to the retrieval of logical operators in infants and children. Taking connectives as the best-known case, we focus first on the little information we have about their preverbal representations (section 1.3.2) and then on the weakly known process of linguistic acquisition. We compare the knowledge of the process of acquisition of linguistic connectives with those of numerals, expressions that may be equivalent to connectives but have a much- better described pattern of acquisition, in order to take advantage of the similitudes between the two and advance in the understanding of connectives (section 1.3.3, but see also chapters 2 and 3 for the experimental development of this idea).

Finally, we concentrate on the generation of sets and its manipulation, since the three remaining subcomponents are deeply related with it. We first present infants' understanding of sets of objects and of situations, to conclude with a brief discussion on what we know about children's abilities to order sets of concepts via entailment (section 1.3.4, but chapters 4 and 5 will then exploit this knowledge to extend our understanding of its development).

1.1 The relationship between language and thought

Human conceptual abilities have captivated philosophers since philosophy exists. How could something as immaterial as minds or thought be studied? What does it consist of? For some, part of the answer to this question can come from 'logic', the study of the valid rules of inferences and consistency of beliefs (Hodges, 2001). Logic and formal logical systems have been used, since Aristotle, to describe coherency in thoughts.

More often than not, thought was discussed in relation to another indubitable human trait: language. For our purposes, we describe language as a computational system of rules and signs that are combined to form complex expressions, and that has among its functions that of transmitting thoughts to communicate mental states.

The relation between the two and, concretely, the amount of conceptual resources they share has always been a topic of controversies. Language could be seen as the externalization of thought, the link between 'things in our mind' and the outside world; or, looking at the other side of the coin, thought could be the internalization of language, something that appears only when we are already proficient speakers.

The significance of this debate is larger than the simple chicken-and-egg problem of understanding which one comes first. Considering language as the externalization of thought implies the existence of powerful thought operations to guide language acquisition and probably the whole cognitive development: under this perspective, it is the ability to entertain thoughts, and more specifically hypotheses, that shape and guide development and, maybe, behavior from the very beginning.

Alternatively, considering thought as an internalization of language means giving more power to the environment in which a child is immersed. It implies that the ability to generate abstract thought does not exist, or is severely reduced, until we have acquired words; potentially, it also implies that to conceive a thought depends on our mother tongue – the famous Sapir-Whorf hypothesis (Boroditsky, 2006). Thus, the relation between language and thought is not a trivial question, but one on which the very nature of human knowledge depends.

Debates around this relationship go back to the ancient Greece. Plato, in Cratylus, argued for a prevalence of thought over language (Sedley, 2018); while sophists defended the inverse vision (Barbosa, 2015). The question remained unsolved over the history of philosophy, and still is in our days.

Among those who defended a substantially Platonic view is Gottlob Frege. His ideas have deeply influenced contemporary philosophy and the debate we are considering. According to the German logician, thoughts are just descriptions of two objects, the True and the False. Therefore, logic, the discipline concerned with the rules of Truth, is at the basis of thought. This approach allowed him to develop a method of formally representing the logic of thoughts and inferences, what we would now call the predicate calculus (Zalta, 2020).

Interestingly, and like Plato proposed, Frege argued that the structure of thought does not come entirely from the outside world but is something ‘from the inner world’, which cannot not be learned by experience since *experience itself* is filtered by it (Frege, 1918).

While thought, in these theories, was structured logically, language was far messier, filled with contradictions and paradoxes. The relation between the two was partial: sentences can express thought but not every sentence does. Conversely, not every thought is encoded by a sentence (Frege, 1918). Thus, Frege, painted a clear picture: thought is logically structured, exists independently of the thinker, and has a limited relationship with the learned, less structured, natural language.

His ideas remained confined in the field of philosophy, while psychology became widely dominated by behaviorism. This diminished the role of innate operations and structures, negating the ontological reality of perceptual and cognitive processes (Fodor, 1985b). In this framework, mental content was to be eliminated, reduced to a manifestation of conditioned reflexes and discriminative responses.

Other views of the relationship between language and thought

Not by chance, the prevalence of behaviorism in psychology came also with the birth of linguistic determinism. Quinean linguistic determinism understands adult representations of the world as a cultural construct (learned through natural language), and, as a such, a construct that can change from generation to generation and from culture to culture (Harman, 2011).
Quine's ideas are deeply related to the Sapir-Whorf hypothesis: each language enshrines a way of perceiving, analyzing, and acting in the world. That is, insofar as languages differ so do speakers' thoughts (Boroditsky, 2006). Linguistic determinism was a strong version of this hypothesis, nowadays mostly dismissed, according to which thought and action are entirely determined by language. A milder way of that hypothesis, linguistic relativity, still widely diffused, proposes that language *shapes* thought (Boroditsky & Ramscar, 2002). Nowadays, linguistic relativity is often discussed within a wider theory of cognition, *embodied cognition* (Boroditsky & Ramscar, 2002), which sees semantic and conceptual information as based on information materialized in action and perceptual systems of the brain (Pulvermüller, 2013). For that view, abstract domains and concepts, which are far less general and far more situation-based, are understood through analogical extensions from more experience-based domains (Pulvermüller, 2013).

In parallel, but completely opposed to behaviorism, the so-called Cognitive Science revolution (Nadel & Piattelli-Palmarini, 2003) brought Frege's ideas back to the center of debate again.

One of the most influential names within the new current was Jerry Fodor. In contraposition to behaviorism, Fodor's theory proposed the existence of 'internal properties' which mediate between inputs and outputs (Rives, 2020). These internal properties are belief-like states, representing the world and guiding behavior; and desire-like states, representing one's goals and motivating behavior (Fodor, 1985a).

Fodor thought of belief-like and desire-like states as abstract, computational, elements, not tied to any specific physical system (Rives, 2020) although realized in them. In a *psychologized* version of Frege's idea, Fodor thought of mental states as possessing constituent structure and intentionality -that is, as carriers of content, as semantically evaluable objects with satisfaction conditions. He argued that thoughts are structured in a *Language of Thought* (Fodor, 1989). The fact that thought has constituent structure, Fodor argued, follows from it being expressed by a highly structured object -language-, and from its systematically -once one can think that ‘John loves Mary’, also thinking that ‘Mary loves John’ is automatically possible-. These two characteristics can only be obtained if we assume that thoughts are composed in terms of structured constituents (Fodor, 1989). Because of this conception, also the ‘logicality’ of thought, rather its their ‘embodiness’, came back to the fore, and with a vengeance.

Fodor’s ideas, like to those of Frege, delineate an innate, highly structured, conceptual repertoire [1], interconnected to language by means of logical structure. Various philosophers besides Fodor assumed that some form of mental logic stands at the basis of thought. These theses connect the structure of thought to the structure of language (Braine & O’Brien, 1998). That view has an implication for language: because it is a reflection of thought, language must have as much logic as thought does. Therefore, this way to conceive the relation between thought and language also opens a way to reverse-access the logicality of thought: via the logicality of language.

But, is language *logical* at all?

1.2 The logicality of language

1.2.1 Montague and the birth of modern formal semantics

To a certain extent, a measure of structure in language has been acknowledged since ancient times (Bod, 2013). However, concessions to structure mostly concerned syntax; to consider that the meaning it conveys is likewise structured is a more recent development.

Tied to Plato’s observation that language allows one to talk falsely (Sedley, 2018), and influenced by the neopositivist point of view, many philosophers disregarded natural language as the best source to express meaning, opting instead for regimented artificial languages born to make mathematical and philosophical argumentation more transparent (Carnap, 1968). Indeed, so did Frege (1879) himself, who never applied his begriffsschrift to reconstruct ordinary reasoning expressed

[1] Fodor’s nativist position even defended the thesis that all mental (non-compositional) concepts are innate. Having innate mental concepts, he argued, would give a clear evolutionary advantage in freeing infants to have to ‘reinvent the wheel’ and learning about human common-sense psychology before learning cultural and knowledge traits (Fodor, 1987)

in natural language, or Russell (1905), who concluded that 'language obscures the view or real meaning'.

The shift of paradigm came with Montague. Creating a real revolution in the field of semantics, he argued that even the semantics (and syntax) of ordinary natural languages could be described in mathematical terms, just as if a natural language was as logical as an artificial language (Montague, 1970).

Montague put the notions of 'true sentence' (under a given interpretation) and entailment to the center of linguistic research (Montague, 1970). Moreover, the formal approach proposed by Montague naturally brought him to study how the meaning of a whole sentence depends on the meaning of its parts; Montague showed that semantic compositionality could be preserved by following how syntax combines expressions in a natural language (Partee, 1984). Crucial to his theory was the idea that there is a systematic relationship between syntax and semantics (Montague, 1970).

In Montague's approach, heavily dependent on an application of set theory to natural language, some operators (e.g., 'most') were treated as higher-order logic operators quantifying over sets of sets. This jump to second-order logic also led him to elaborate a general treatment of natural language expressions such as noun phrases as generalized quantifiers[2]. With this treatment, what first-order logic would treat as expressions of separate category, such as nouns ('John') and quantifier-composed expressions ('Most students'), turn out to be in the same category, and can be composed by applying the same functional composition that can be found in English syntax. Thus in his approach, most differences between syntax and semantics in the description of linguistic meaning are leveled, and from it the logical structure of even plain natural language sentences surfaces (Janssen, 2020).

His theory, although very powerful, was meant to be descriptive, and did not seek to have any psychological grounding (Partee, 2004). For him, linguistics was to be seen as a branch of mathematics.

After his death, the field that he helped to create, formal semantics, experienced great growth and became soon influenced by Chomsky (1980)'s ideas. While keeping faithful to Montague's claim that that natural languages can be generated by formal systems, it took from the generative tradition the idea that such representations constitute accurate models of the implicit knowledge underlying language competence. To put it in another way, this point of view entails that the descrip-

[2] In a a generalized quantifier treatment, the denotation of any descriptive phrase is a set of properties; for example 'John' would mean 'the set of properties that hold for him'.

tions provided by the theory are not just mathematical artefacts, but have some degree, perhaps extreme, of psychological reality (Chierchia & McConnell-Ginet, 2000).

1.2.2 The borders of semantics

The renewed interest in psychologically-valid representations of natural language meaning had the important consequence of bringing to the fore a theme largely overlooked: the study of the interactions between semantics and pragmatics (Chierchia & McConnell-Ginet, 2000). Unlike formal semantics of natural language, which had become a formal and well-defined theory since Montague and before, traditionally pragmatics played for reasoning theories the role that lexicon played in linguistics: everything that was idiosyncratic, that resisted systematic treatment, or that fell into cross-cultural and cross-linguistic variability was judged to be pragmatics.

One of the paradigmatic case studies at the border of semantics and pragmatics rapidly became that of words with direct logical correspondence, which we now discuss.

Words with logical valence and scalar implicatures

Words with direct logical correspondence are words such as quantifiers (‘some’, ‘all’), numerals (‘one’, ‘two’ which can be seen as ‘exact quantification’) and connectives (‘and’, ‘or’). The connection between them and standard logic is so strong that usually quantifiers are semantically analyzed as quantifications over a domain (Fox, 2007; Chierchia, 2013), just as in standard logic. Even fine-grained differences between natural language quantifiers, such as the difference between ’all’ and ‘each’ (or collective and distributive quantifier) have been directly tied to properties of standard logic; specifically, they have been claimed to be mentally represented as second and first-order logic operators, respectively (Knowlton, Pietroski, Halberda, & Lidz, 2019). Also (relatively) simpler connectives such as ‘and’ and ‘or’ have traditionally been assumed to be equivalent to the logical $\vee$ and $\wedge$ (Beilin & Lust, 1975a).

For our purposes, it is interesting to notice that in all these cases of paradigmatic relationship between logic and language, along with an *at least*, roughly logical meaning, these expressions have a second, more enriched reading (example (2) represented here as (4)):

(3) If you open the door, I'll take some animals outside.

a. ***At least* meaning**: I'll take more than one animal (possibly all) outside.

b. **Enriched meaning**: I'll take more than one animal, but not all, outside.

(4) If you open the door, I'll take the cat or the dog outside.

a. ***At least* meaning**: I'll take the cat or I'll take the dog or I'll take both animals outside.

b. **Enriched meaning**: I'll take the cat or I'll take the dog but I won't take both animals outside.

In everyday language, these words are usually interpreted according to this second meaning. Nevertheless, such a meaning appears to be processed more slowly (Bott & Noveck, 2004; Breheny, Katsos, & Williams, 2006; Huang & Snedeker, 2009; Tomlinson Jr, Bailey, & Bott, 2013) and is apparently difficult for children to understand (Paris, 1973; Braine & Rumain, 1981; Chierchia et al., 2001; Guasti et al., 2005; Barner et al., 2011). These facts, along with the convergence of the *at least* meaning with classical logic, have been taken as evidence that the enriched reading is derived from the *at least*, logical meaning via some kind of process or conversational shorthand (Stiller, Goodman, & Frank, 2011).

Such process, aimed at restricting the meaning of the logical valence word, is called 'Scalar Implicature'. An impressive amount of research over the last 40 years has tried understand the nature of scalar implicatures, as they are seen as the ideal intersection between semantics and pragmatics.

The most common description of scalar implicatures appeals to the Gricean maxim of quantity [3].

[3] Indeed, sometimes they are even called *Quantity-based implicatures* (Noveck, 2001; Horn, 2006)

Grice (1975) standard conversational norms

Grice created a list of conversational rules which are supposed to be the basis of how utterances get their meanings in conversational exchanges. Clearly, they put much of the weight of the interpretation onto pragmatic competence. The maxims are reported below:

1. **Quantity:** make your contribution as informative as required but do NOT make it MORE informative than required.

2. **Quality:** tell the truth and avoid statements for which there is insufficient evidence.

3. **Relation:** be relevant.

4. **Manner:** avoid ambiguity, confusion, and obscurity.

As an example, the derivation of an enriched meaning for a disjunctive expression by means of the Gricean maxims would proceed as follows (adapted from Chierchia (2017):

(5) The speaker said '(If you open the door) I'll take the cat or the dog outside'.
  a. The speaker did not say 'I'll take the cat and the dog outside'.
  b. We assume that 'I'll take the cat and the dog outside' is relevant in this context
  c. We take into account the maxim of quantity: 'Say only things which you believe to be true'.
  d. We take into account the maxim of quality: 'Say *all* relevant things which you believe to be true'.
  e. By the maxims of quantity and quality, if the speaker believed that she would take the cat and the dog outside, she would have said so.
  f. Therefore, the speaker does NOT believe that she would take both the cat and the dog outside.

This process, illustrated in (5), can also be described as one of considering relevant alternative sentences (in the previous case, 'I'll take the cat and the dog outside') and negating them to obtain the enriched meaning (Chierchia, 2013). However, which sentence is relevant, and which one is not, is not a trivial question. The traditional Gricean view proposes to draw alternatives from Horn scales.

Horn scales

A Horn scale is a list of semantically-related expressions, sorted from the more to the less informative one (Horn, 1989). [a] Some examples of such scales, which would provide the relevant alternatives to give rise to scalar implicatures, are as follows:

1. Quantifiers:<all, many, some>[b]
2. Connectives:<and, or>
3. Cardinal number terms: <..., three, two, one>
4. Modals: <must, might>

[a] Understanding what 'informativeness' means raises subtle issues. Here we can treat it as some sort of 'situational restriction'. An expression is more restricting and thus, more informative, if it can be applied to less situations. For example, the word 'dog' is more restricting and more informative than the word 'animal', since it can be used in less cases than 'animal': while we can say 'I saw an animal' when in view of any dog, the inverse is not true. In a similar way, a word such as 'all' is also more informative than 'some', since 'some' can be applied to situations in which we could also use 'all', but the contrary is false. Horn scales are also usually explained in terms of entailment, see section 1.2.2 for further details on this issue.

[b] We will use the notation '$< a, b, c >$' to refer to ordered sets, that is sets (i.e- *groups*) that follow a specific order. For example, in that case, <all, many, some>, the words are ordered by the most informative ('all'), to the less informative ('some').

At the same time, Horn scales are not the only source of alternatives for scalar -like implicatures: these can also come from the context. For instance, if the context involves two people with glasses, it is possible to construct scales from ad-hoc concepts, such as 'the friend with glasses', as opposed to 'the friend with glasses and a hat', where the first sentence will identify the person with *only* glasses (Stiller et al., 2011). Interestingly, Chierchia (2017) noticed that not all concepts qualify to create a scale, but that scalemates must be *monotonically consistent*, that is, they must have the same increasing or decreasing value. For example, a scale including <some, just some> would not be acceptable because it does not have a uniform monotonic construction: 'some' is upward monotonic (that is, allowing inferences from less to more informative, from 'some' to 'all') ; while 'just some' is nonmonotonic (it does not allow inferences either from less to more informative or from more to less informative. See also 1.2.2).

It is important to notice therefore, that this process, thought to be unfolded as a pragmatic reasoning (e.g., by applying Grice's maxim of quantity), requires the intervention of a logical/mathematical concept: that of the monotonicity of the relevant alternatives.

Entailment relations and their role in meaning

The pervasive presence of logic in the process, however, does not stop here. A very well-known fact regarding scalar implicatures is that their existence tends to depend on the logical linguistic context in which the scalar word is embedded (Horn, 1989; Chierchia et al., 2001; Chierchia, 2013). That is, some logical linguistic environments favor the emergence of such implicatures, while others disfavor it. The key feature determining whether or not scalar implicatures may appear is the different pattern of entailment of the grammatical context.

Entailment relations and set theory

Entailment is a relationship between two sentences such that if the first is true, the second will necessarily be true (Chierchia & McConnell-Ginet, 2000). To give an example particularly relevant for us, the sentence

(6) Mary eats pizza and cake

entails that:

(7) Mary eats pizza

because in all situations in which the first sentence is true, the second one will also be. This notion, and the spontaneity of such inference (which we will investigate in chapter 4), indicates that knowledge is organized in sets, supersets, and subsets.

For example, one could easily imagine a set of ‘dogs’ comprising several dog breeds (dalmatians, huskies, terriers, bulldogs...). In that context, each of the dogs’ breed would be a *subset* of the set of the *superset* of dogs. At the same time, it is not hard to imagine our set of dogs to be part of a bigger set, for example, of mammals. In that context, the set of dogs would be a *subset* of ‘mammals’. Alternatively, the set of ‘mammals’ would be the *superset* of the set of ‘dogs’.

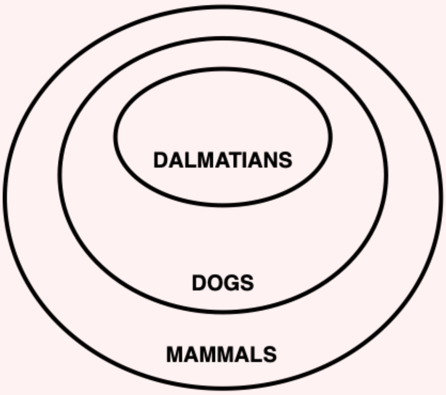


Figure 1.1: Structure of the set of Dalmatians, included in the set of dogs, included in the set of mammals

We end up, therefore, constructing a special case of a lattice of concepts, in which supersets include subsets, and are included by wider supersets. Ontological issues aside, this structure is useful to expand our knowledge. For instance, knowing about the set of ‘mammals’ that they are ‘vertebrate animals that produce milk’ leads us to safely infer that ‘dogs’ are also vertebrate animals that produce milk. Since in all situations in which mammals are ‘vertebrate animals that produce milk’, ‘dogs’ will also be vertebrate animals producing milk; we can conclude that the first sentence, about mammals, entails the second sentence about dogs.

Contexts in which an expression in a sentence entails the truth of that sentence in all its supersets are called Upward entailing contexts (UE). To give a juicer example, the sentence:

(8) If Mary goes inside the tent, she eats pizza.

creates an upward entailing context, because one can draw the inference towards the superset (food) and conclude that ‘If Mary goes inside the tent, she eats food’. Notice, at the same time, that the same inference cannot be drawn towards the subset —the type of pizza eaten— because from the truth of (8) one cannot infer that if Mary goes inside the tent, she will eat pizza with olives.

By contrast, contexts in which the expression entails the truth of the sentence in all its subsets are called downward entailing (DE). For example:

(9) If Mary eats pizza, she gets stains in her t-shirt.

The preceding example (9) leads us to infer about the subset and conclude that 'If Mary eats pizza with olives' she will get stains. In contrast to the previous example, however, we cannot conclude anything about the superset, – nothing about whether Mary gets stains when she is eating food or not.
Interestingly, Downward Entailing contexts reverse the informativeness of Horn scales: 'If Mary eats two slices of pizza she gets stains' is less informative than 'If Mary eats one slice of pizza she gets stains', since, in that context, 'eating one slice' entails 'eating two slices', but not vice versa.

Scalar implicatures are typically licensed in Upward Entailing Contexts, while they are not drawn (or they are 'canceled', depending on the theory) in Downward Entailing Contexts. The pulse for this division is quite strong, as it can be seen from the following examples, adapted from Chierchia (2013):

	Upward entailing contexts: Enriched interpretation	**Downward entailing contexts: 'at least' interpretation**
Conditional (antecedent vs. consequent)	If Mary goes inside the tent, she eats pizza or cake.	If Mary eats pizza or cake, she gets stains in her t-shirt.
Episodic vs. generic sentences	A Dalmatian or a Husky barked yesterday night.	A Dalmatian or a Husky always barks at nights.
Positive vs negative sentences	Mary intends to eat pizza or cake.	Mary does not intend to eat pizza or cake.
Scope of positive vs negative quantifiers	Somebody in the building has either a Dalmatian or a Husky.	Nobody in the building has either a Dalmatian or a Husky.

Table 1.1: Comparison of contexts in which scalar items typically get an enriched or an 'at least' interpretation'

If we take a context in which the expression is embedded inside a conditional (first cell of the table 1.1), it is not hard to imagine that Mary will eat one or the other food. However, if we exchange antecedent and consequent as in the second cell of the first row then the sentence implies that taking *any of the two* will make Mary dirty. The difference is also evident in the second context: an episodic sentence like 'A Dalmatian or a Husky barked yesterday night' strongly suggests that one of them barked, but not both. However, if you uttered 'A Dalmatian or a Husky always barks at nights' you should be prepared to have hard nights if you decided to adopt any one of the two dogs, and even more so if you adopt both.

The division of contexts according to logical entailment is relevant not only in the derivation of scalar implicatures but also for several other linguistic phenomena.

Consider, as another striking example, negative polarity items (NPI) [4] , like *any* and *ever*. These items are only permitted in DE contexts (Chierchia et al., 2001; Chierchia, 2013), as table 1.2 shows.

	Upward entailing contexts: agramatical NPIs	**Downward entailing contexts: gramatical NPIs**
Conditional (antecedent vs. consequent)	*If the doctor understands, I will ever have a slice of pizza	If I ever have a slice of pizza, the doctor will understand
Episodic vs. generic sentences	*Any dog barked yesterday night	Any dog always barks at nights
Positive vs negative sentences	*Mary intends to ever eat pizza.	Mary does not intend to ever eat pizza.
Scope of positive vs negative quantifiers	*Somebody in the building has any dog	Nobody in the building has any dog

Table 1.2: Comparison of contexts in which NPIs are grammatical

The similitudes between the behavior of scalar implicatures and negative polarity items are very evident. It is clear that their interpretation and grammaticality depend on the form and on the embedding patterns in which an expression resides: semantic factors with a direct correspondence into the syntax of these expressions are fundamental. These considerations point at an even vaster and more pervasive role of logic in language. Indeed, Fox (2007), Chierchia (2013) or Singh, Wexler, Astle, Kamawar, and Fox (2013) developed theories to treat the above mentioned phenomena in a unitary way. Such a treatment recognizes a fundamental role to logic in structuring both the language and the human minds that use it. As Chierchia elaborated the most thoroughly developed proposal along these lines (Chierchia, 2013, 2017), we will now present it in more details.

1.2.3 The theory of exhaustification

In Chierchia's account, the operation of exhaustification has such a fundamental role that we will refer to this account with the well-deserved name of *theory of exhaustification*. Chierchia (2013) defines exhaustification as an operation that, given a sentence, applies to a set of relevant alternatives. It negates them and in

[4] Negative Polarity Items are lexical items that are only grammatical in certain contexts. Traditionally these contexts have been associated to a 'negative' flavor (hence their name).

this way it determines the desired meaning of the targeted sentence. These alternatives differ in the quantity of information they convey and are ordered on a scale, in terms of logical strength. The process, equivalent to the one we described in (5), differs from the Gricean explanation in that exhaustification is a semantic phenomenon marked in the grammar. While the Gricean maxims work as pragmatic principles, which apply after the semantic meaning has been retrieved, exhaustification is a logical operation that applies according to an established set of rules which depend on the meaning of the words and on the logical context into which they are embedded.

Exhaustification requires a silent operator (O), which is assumed to be present in the compositional system (in a similar way as silent operators that have been proposed for the null subject (pro), see Chomsky (1993)). O functions as a kind of 'silent only'. It 'focuses' on the uttered sentence and negates ('exhaustifies') the unuttered ones. This behavior is alike that of the explicit 'only':

(10) Only Mary ate pizza with olives

In example (10) 'only' rules out the possibility that other people, John and Laura, for example, also ate pizza with olives. Similarly, the silent operator would target the word with logical valence and enrich the meaning of the sentence ensuring that the asserted sentence is the only true member of the relevant set of alternatives:

(11) Mary has a dog or a cat.

a. **Logical interpretation:** [Mary has a dog ∨ Mary has a cat]

i. Logical parse: $[D \vee C]$ (where D stands for *Mary has a dog* and C stands for *Mary has a cat*)

ii. Reading: Mary has a dog or a cat (or both).

b. **Enriched interpretation:** O[Mary has a dog ∨ Mary has a cat]

i. Relevant alternatives: $[D \wedge C]$

ii. : Logical parse: $O_{\underline{[D \wedge C]}}[D \vee C] = D \vee C \wedge \neg(D \wedge C)$

iii. Reading: Mary has a dog or a cat but not both.

How alternatives are selected is a crucial part of the theory. Alternatives may be lexically determined (for example via Horn scales, see section 1.2.2) and modified via the monotonicity constraint, or else be associated with domain alternatives. An example of domain alternatives is given in (10): in that case, 'only' parses out alternatives given by the domain ('the (extralinguistic) situation'): other people (for example John and Laura) are excluded.

The evaluation of the quantity of information that alternatives bear, which is a fundamental prerequisite for its ordering, is a process reminiscent of the Mathematical Theory of Communication (MTC). In both theories, the amount of the information carried by a sentence (or situation) depends on the number of situations (or, in formal semantics terminology, the number of 'possible worlds') to which the sentence can be applied(Floridi, 2019). The larger the number of situations to which it applies, the lower its degree of informativeness. Therefore, a sentence like (12) is more informative than (13). In the same way, (15) is more informative than (14), since it can be applied to fewer situations.

(12) Mary has a Dalmata.

(13) Mary has a dog.

(14) Mary ate some pizza yesterday (in its logical sense: 'at least some').

(15) Mary ate all pizza yesterday.

Alternatives may or may not be active depending on the specifics of the word with logical valence and the relation it establishes with the operator. However, once alternatives are active, the logical context of the sentence plays a crucial role in their exhaustification. The reason for that has to do with the fact that exhaustification takes into account the informativeness of alternatives: only logically stronger, more informative alternatives are negated. As we anticipated in section 1.2.2, and we commented above, stronger alternatives are those that are more informative. For instance, example (15) is more informative than example (14) because it can be applied to fewer situations. The fact that only logically stronger alternatives are negated explains the asymmetrical distribution of Scalar Implicatures as well as why Negative Polarity Items are only grammatical in Downward Entailing contexts:

(16) If you open the door, I'll take the cat or the dog out.

a. Alternative: I'll take the cat *and* the dog out.

b. $and \vDash or$('and' entails 'or'), therefore 'and' is logically stronger

c. So we negate the alternative with 'and' and in the main sentence, 'or' gets an enriched meaning

Example (16) shows the process of exhaustification for the sentence 'If you open the door, I'll take the cat the a dog out'. Since the logical context where the word with logical valence ('or') is upward entailing (that is, licensing inferences from the set to the superset, see section 1.2.2 for more details), the alternative sentence 'If you open the door, I'll take the cat and the dog out' is logically stronger (it can be applied in less situations than our target sentence). Therefore, it is negated by the operation and we read the 'or' in the target sentence 'If you open the door, I'll take the cat and the dog out' as having an enriched meaning (the cat or the dog but not both). In contrast, consider what happens in example (17):

(17) If you take the cat or the dog outside, I'll close the door

a. Alternative: You take the cat **and** the dog outside.

b. In that context, *and* $\nvDash$ *or* ('and' does NOT entail 'or'), therefore 'and' is not logically stronger

c. So we do *not* negate 'and' and 'or' gets an *at least* meaning.

In this example, the context in which 'or' is embedded is downward-entailing (licensing inferences from the set to the subset). That means that the alternative sentence 'If you take the cat and the dog outside, I'll close the door' is not logically stronger, and so it cannot be ruled out by exhaustification. Therefore, 'or' in our target sentence keeps its *at least* reading ('the cat or the dog, maybe both').

As (16) and (17) illustrate, in a context with a word with logical valence ('or' in that case), alternatives are evaluated in terms of logical strength. Alternatives entailed by the uttered sentence (that is, logically weaker, as in (17)) are not activated and thus not considered, only those which are stronger contribute to the enriched meaning.

If we focus on 'any', a Negative Polarity Item, applying the same reasoning helps us to see why it is only grammatical in Downward entailing contexts:

(18) Is there some slice of pizza left?

(19) There isn't any slice of pizza left.

a. Alternative: There isn't a slice of pizza [5]

[5] Following (Chierchia, 2013), we assume that existential statements like 'There isn't a slice of pizza left' come with a contextually supplied domain variable (similar to a predicate) that determines the range of the quantifier. In that characterization, 'any' is akin to an existential item like 'a', the difference being that its domain is broader (i.e. it includes more specimens of the general kind, such as 'frozen pizza', or 'pizza stored in unusual places'), than that of 'a'. Alternatives of 'any' are drawn from this difference in domain width (including as alternatives smaller domains), and not from

b. The alternative is not logically stronger than ‘any’ (since ‘any’ $\approx$ ‘some’, and ‘There isn’t a slice of pizza’ $\nvDash$ ‘There aren’t some slices of pizza’)

Example (19), a downward entailing context, works in a similar way as we saw in example (17): it recruits the alternatives, but since they are not logically stronger, it cannot negate them.

However, if we insert ‘any’ in an upward entailing context, we encounter a contradiction:

(20) *There is any slice of pizza left.

a. Alternative: There is a slice of pizza

b. The alternative is logically stronger than ‘any’ (since ‘any’ $\approx$ ‘some’, and ‘There is a slice of pizza’ $\vDash$ ‘There are some slices of pizza’).

c. So we have to negate it and the meaning becomes ‘There is some slice of pizza left and there is not a slice of pizza left’.

Example (20) follows the same procedure as example (19), but since in this context the alternatives are logically stronger than our target sentence, they have to be negated; which results in a contradiction (the meaning of ‘there is any slice of pizza left’ would be something similar to: ‘there are some slices of pizza left and there is no slice of pizza left’).

Examples (19) and (20) show a general moral: exhaustification, a semantic/logical process, may percolate up to grammar, fixing legal and illegal uses of lexical items with logical content. Indeed, for Chierchia, all negative polarity items are a grammaticalization of the need to exhaustify via the construction of alternatives.

Thus, the broader picture that appears from the theory should begin being clearer: traditional approaches to several phenomena – among them, Negative Polarity, and Scalar Implicatures – often resorted to ad-hoc explanations; by contrast, the exhaustification theory unifies (and simplifies) their treatment, by making use of

the actual lexical items. This difference has further consequences, such as explaining the emphasis taste that sentences with ‘any’ hold respective to sentences with ‘a’: borrowing the example from (Chierchia, 2013), ‘I will not vote for any republican’ sounds stronger than ‘I will not vote for a republican’. That would be because the domain related to ‘any republican’ is wider than that of ‘a republican’.

In this example, for sake of simplicity, we encode this difference with the use of the indefinite ‘a’, but one has to keep in mind that the actual difference is not related with the choice of the word, as in scalar items, but with the width of the associated domain.

a relatively simple operation: exhaustification over a set of alternatives. However, once this perspective is accepted, the role of logic in language has to be accepted as pervasive. This implies that the study of the psychological nature of the logical processes involved becomes a fundamental question to understand language and, with it, thought.

1.2.4 The operational basis of exhaustification

Nowadays the framework of 'exhaustification + alternatives' is accepted by many semanticists (Chierchia et al., 2001; Barner, Libenson, Cheung, & Takasaki, 2009; Barner et al., 2011; Singh et al., 2013). Indeed, other phenomena at the interface between semantics and pragmatics, such as focus (Rooth, 1992), dual morphology (Marušič et al., 2020), free choice (Chierchia, 2013) or the exhaustivity in wh-questions (Schulz & Roeper, 2011) have been explained in a very similar framework.

However, this theory of language does not merely aim to describe how language works, but also how *the mind that produces that language* works (Chomsky, 1993; Chierchia, 2013) . Thus, this framework has to make concrete claims about which mental operations are required in the production and comprehension of the logical components of language.

To try and understand which parts of this theory may have some psychological reality, let us go back again to our pizza example more in details. According to Barner and Bachrach (2010), understanding a sentence like 'Mary likes most types of pizza' would call for the following derivation:

(21) Mary likes most types of pizza

1. Compute the basic (logical) meaning. For example, for 'most', the meaning would be 'more than the half'.
2. Generate a set of alternatives (by replacing the target word with its scalar alternatives, coming from Horn scales (Horn, 1989)). In our 'most' case, these would be 'some', 'many', and 'all'.
3. Order such alternatives in terms of their quantity of information. That requires considering all the alternatives and to compare them to know which ones are more informative.
4. Restrict the alternatives to the set containing the stronger alternatives. In our example, we would drop 'some' and 'many', here, as they are

not logically entailed by ‘most’.

5. Augment (’exhaustify’) the basic meaning by negating all its stronger alternatives. That would mean, for the meaning of ‘most’, to become ‘more than the half, but not all’.

As exemplified in (21), the process requires several steps. First, one has to retrieve and understand the logical meaning of the word. Then, a set of alternatives, drawn from its lexical meaning or the linguistic and extralinguistic context, must be generated. Alternatives then need to be ordered according to informativeness (logical entailment) and the less informative ones (i.e. those entailed by the original word) are dropped. Finally, the remaining set of alternatives is negated to obtain the enriched meaning of the sentence.

Besides the knowledge of the logical meaning of the words, this process seems to require the mastery of a series of set concepts, such as set generation, ordering, set inclusion and exclusion (related with the ability to compute entailment relations[6]) and, finally, the ability to ‘close off’ and eliminate the unnecessary sets in order to restrict the interpretation to the relevant one in the context. A fundamental question for the theory is then to investigate whether, when and how human thought masters these operations, and whether language acquisition influences them.

1.3 The logicality of thought

1.3.1 Logical reasoning and preverbal capacities

A way to address the question is to develop an investigation into the logical abilities of infants and children. Yet, a highly influencing approach still very present in the literature (Piaget & Inhelder, 1951; Xu, 2002, 2019) denies that such an investigation makes sense. For Piaget, the piagetians, and the neo-constructivists, logical reasoning is not within the repertoire of human capacities before seven years, and not before adolescence does it reach it full development, when the stage of formal operations is attained.

Despite the importance of this view in developmental psychology, a growing number of lines of research in infancy seems to suggest that some mastery of logical

[6] The ability to compute entailment relations, even though it draws from set generation, is separate from the former because it relates to the operation of navigating through the set, making potential implications explicit, rather to the structure of the set itself.

operations is available early on and may underpin several cognitive operations. Infants entertain certain kinds of reasoning since they are very young. Before their first year of life, they expect intentional agents to behave rationally, realizing a goal-directed action by using the most efficient trajectory (Gergely and Csibra (2003); they can acquire abstract rules likely involving variables (Marcus, Vijayan, Rao, & Vishton, 1999; Marchetto & Bonatti, 2013; Kabdebon, Pena, Buiatti, & Dehaene-Lambertz, 2015), extract frequency distributions from distributed numerical input (Libertus, Feigenson, & Halberda, 2018) or deal with modality-independent representations of numerosity (Izard, Sann, Spelke, & Streri, 2009; Feigenson, 2011) . All these abilities may be part of infants' core cognitive repertoires: operations that infants know with little or no experience[7]. For some, they could constitute a departing ground for inferential reasoning (Stahl & Feigenson, 2017). There is evidence that infants can take advantage of inferential reasoning: Gweon and Schulz (2011) and Stahl and Feigenson (2015)'s studies suggest that 16 and 11-month-old infants form hypotheses about the world, test them and rule out less likely representations following the evidence they collect. Also 3 to 4 year old children use a similar process of fast acquisition to infer novel meanings of words (Havron, de Carvalho, Fiévet, & Christophe, 2019). Furthermore, there is a striking structural similarity between learning by means of causal and inferential reasoning and exhaustification: just as exhaustification is substantially a process of narrowing down ordered options to the smallest set satisfying some conditions, so is hypothesis testing and confirmation. For instance, in Gweon and Schulz (2011)'s studies, 16 month-old infants were shown a toy that played music when pushing a button. The experimenter then handed the child either the same toy or a differently colored one, and the child unsuccessfully tried to activate it. The critical observation was that when infants had been given the exact same toy that the experimenter was using, they tended to hand the toy to their parents, while when they were given the differently colored toy they tended to reach for the other toy. The authors claimed that infants inferred they were the source of the failure in the exact same object condition, so they sought for help, and they believed that the failure was due to the object in the differently-colored toy condition. This different behavior can be explained by supposing infants created a hypothesis of the possible causes of the malfunctioning of the toy: the agent (themselves) or the world (the toy). When the evidence was compatible with the agent being responsible, infants changed it (i.e. they handed the toy to their parents); and when it was compatible with a problem with the world, they acted in consequence (i.e. they changed the toy).

[7] It may be that such operations are entirely encapsulated inside some specialized module (e.g., naive physics); however, empirical indications suggest that we are dealing with operations available across domains (Cesana-Arlotti, Kovács, & Téglás, 2020)

Both inferential reasoning and exhaustification processes presuppose the generation of a set of relevant alternatives (alternative sentences or alternative explanations), and then their elimination (via estimations of informativeness, in the case of exhaustification, or via their feasibility in the case of hypothesis testing). In short, the possibility exists that the ability to project hypotheses, or to extract general rules, is part of some central system depending, in part at least, on logical resources which may be strongly overlap with those needed for exhaustification.

Beyond this speculative hypothesis about the relation between hypothesis testing and exhaustification, the general study of early reasoning constitute an important entrance gate to the 'thought or language first' question. Unlike adults, infants have not yet started to talk (as the very etymology of 'infant' states, from Latin *in*, 'not', and *fans*, present participle of *for, fari, fatus sum* 'to speak', 'that who does not speak'). Their limited experience with language makes them ideal to understand how language learning interacts with their mental abilities, and until which point the adult reasoning capacities are influenced by verbal behavior or have their origins in preverbal reasoning.

1.3.2 Logical operators in (pre-)verbal minds

One of the most immediate starting points to assess whether infants possess reasoning abilities is to study if an infant mind masters preverbal analogs of logical operators. Work focusing on specific logical operators has sought to understand whether logical functions are available to animal or infant thought, with the final aim to test if they can productively combine with other symbols available in thought – a sign of compositionality, a fundamental property of semantics and of any 'language of thought' (see section 1.2.1).

In recent years, research has intensified but results have been mixed. For instance, work on the preverbal understanding of logical negation (truth-functional negation) suggests that infants may not possess a full negation before two years of age (Feiman, Mody, Sanborn, and Carey, 2017, but see De Carvalho, Dautriche, Christophe, and Trueswell, 2019 for a different opinion). Even the possibility to conceive an apparently simple concept related to negation, such as the concept of 'different', appears to be beyond the repertoire of conceptual primitives (Hochmann, Mody, & Carey, 2016; Hochmann, Carey, & Mehler, 2018).

However, more solid evidence of early availability exists for one of the best-studied cases of elementary reasoning: disjunctive syllogism (A or B, not A; therefore B).

The interest in studying this pattern is multiple [8]. First, it could be a basic tool to handle reasoning with multiple hypotheses; second, it offers a minimal test case for some degree of compositionality. To understand the disjunctive syllogism, one needs: (1) to represent a disjunction between (at least) two possible states of affairs, (2) to have a way to eliminate one piece of information, perhaps with some form of negation (or full negation according to Carey (2009)), and (3) to combine them to generate novel information. Thus, in this process of reasoning, merging together hypothesis formation and elimination on the basis of a preverbal representation of a disjunction seem involved.

Research has especially focused on disjunctive syllogism in children. A substantial number of studies has come to the conclusion that only children older than 2.5 years (that is, not infants and not animals) make use of disjunctive syllogism (Mody & Carey, 2016). Other evidence, however, is difficult to square with this conclusion. In particular, the finding that 12 month-old infants draw conclusions by some form of elimination of alternatives (Cesana-Arlotti et al., 2018), and use those conclusions to make further inferences in other domains (Ekramnia, 2016; Cesana-Arlotti et al., 2020) paint a much more complex picture which deserves being investigated.

1.3.3 Logical operators and numerals in children

One way to pursue this investigation further, albeit in a more indirect way, is to understand the development of words encoding logical operators. 'Or', one of the focuses of our investigation, starts being produced shortly after 'and', at around 3-3.5 years of age (Morris, 2008). We do not know much of what happens before, but by that time, children have no problem distinguishing between the meaning of 'and' and 'or' (Jasbi & Frank, 2017; Pagliarini, Crain, & Guasti, 2018). Even so, they seem to struggle to understand the enriched meaning of 'or' (see chapter 1.2.2) until they are around 5 years of age (Braine & Rumain, 1981; Chierchia et al., 2001; Jasbi & Frank, 2017). At the same age, children also appear to have difficulties to understand the enriched meaning of other words with logical valence, such as quantifiers (Noveck, 2001; Pouscoulous, Noveck, Politzer, & Bastide, 2007; Barner et al., 2011; Foppolo et al., 2012) . Our very limited

[8] Interestingly, although perhaps more anecdotical, also this form of reasoning has some resemblances with exhaustification, in a similar way to that presented before for the process of testing hypotheses: it requires representing a set of alternatives (for the basic case of a disjunctive syllogism, only two) and ruling them out (in the basis of new information, for the disjunctive syllogism, in the basis of informativeness for exhaustification).

knowledge of how words with logical valence are acquired contrasts with the well-documented development of another category of logical words intimately tied to them: numerals. The acquisition of numerals is a compelling example of how mental concepts are sequentially and progressively mapped to linguistic words. Infants possess several sources of elementary numerical knowledge. An innate system of approximate number (called ANS, Approximate Number System) allows them to distinguish between quantities in a certain ratios, but not below it (Feigenson, 2011). Another system, *parallel individuation* (Carey, 2009; Carey & Barner, 2019), represents each object by a unique mental symbol, spatiotemporally determined (Carey, 2009). Although precise, it has a limit of object-based attention of around three objects in infancy, and four in adults (Le Corre & Carey, 2007; Zosh, Halberda, & Feigenson, 2011). As ANS, this system is evolutionary primitive, as it has also been shown to exist in several other species (Barner, Wood, Hauser, & Carey, 2008). Using parallel individuation, infants can spontaneously represent up to exactly three objects in parallel, but still fail to represent 'exactly five', or 'twenty-four'. In isolation, neither the approximate nor the exact systems are the basis of the natural number system or calculus (Le Corre & Carey, 2007; Barner, 2017). Some extra ingredient is needed. For numerals, what is needed is to learn to map the pre-existing representations given by parallel individuation and number words, as well as a process of bootstrapping the recursive nature of the numerical series (Chu, Cheung, Schneider, Sullivan, & Barner, 2020; Schneider, Sullivan, Guo, & Barner, 2021). This process of acquisition starts by learning the counting words around the second birthday: children learn to recite numbers as a routine, without tying them to their meaning (Wynn, 1992). Shortly after that, they start a sequential mapping process from each number word (first 'one', after that 'two') to the corresponding quantities. This effortful, item-based process (Wagner, Kimura, Cheung, & Barner, 2015; Schneider, Pankonin, Schachner, & Barner, 2021) culminates when children reach the limits of their parallel individuation system and map the word 'three' or 'four' to their reference. Only then they generalize to all the numerals from their counting list.

The difference between the understanding of the acquisition of numerals and that of connectives is puzzling if we consider the unitary theoretical treatment that words with logical valence are given in formal semantics, all of them being words with logical valence (numerals can be seen as 'exact quantification'), showing a duplicity of meaning (*at least* and enriched meanings) and subject to exhaustification [9] (Chierchia, 2013). Such uniformity would lead one to expect some paral-

[9] There are some differences as well, since the 'enriched' meaning of numerals seems to be mastered earlier by children than the enriched meanings of quantifiers or 'or' (see for example Huang, Spelke, and Snedeker, 2013). Chapter 2 delves deeper into the relationship between numerals and

lelism in the development of these concepts.

1.3.4 Generation of sets and estimations of informativeness

Besides the mastering of logical operators, another important piece is needed for exhaustification: the understanding of set concepts (set inclusion and exclusion) and ordering of alternatives according to their informativeness. Can preverbal infants deal with set structures?

Sets appear to be fundamental for the phenomenon of *chunking* in memory. Just like adults, even infants as young as fourteen-months can exceed the limits of their working memory (set at three objects tracked in parallel) by chunking objects into sets, based on their location or property/kind information (Feigenson & Halberda, 2004) or labels (Feigenson & Halberda, 2008). In that way, infants (and adults) are able to maintain the representation of more than three objects: they can maintain the chunk representations *containing* the objects.

In each of these chunks, for infants, two individual objects seem to be the maximum number that can be bound (Feigenson & Halberda, 2008). However, chunks can be re-chunked to create 'superchunks', or, in set-theory terminology, 'supersets': that iterative process shapes a hierarchical-structured memory with at least three levels of memory representations, which can encode up to eight objects, vastly overcoming their working memory limit (Rosenberg & Feigenson, 2013; Zosh et al., 2011). It seems that this process is reasonably automatic: once the appropriate structure is given, infants cannot 'ignore' the superset: this must be computed and encoded in memory (Zosh et al., 2011). Thus, human memory seems to make use of set structures from an early age.

While the representation of several sets of objects in parallel seems well within the reach of infants, there are more serious doubts as to whether the representation of 'sets of possibilities', multiple possibilities in parallel, is possible at a young age. Besides being a special case of set representation, the representation of multiple possibilities is important for our inquiry because they constitute a non-linguistic equivalent to alternatives in the exhaustification process: they are alternative hypotheses of how a situation could end. Studies demonstrate that from the age of 3 children can represent multiple exclusive versions of the future, but what happens before is more debated: Redshaw and Suddendorf (2016) defend that younger infants do not have that capacity, whereas Téglás, Girotto, Gonzalez, and Bonatti

connectives.

(2007), Téglás and Bonatti (2016) propose some form of multiple exclusive possible outcome representations in preverbal infants.

The ability to represent possibilities is often investigated in relation to the ability to track probabilities, which is directly related to the estimation of informativeness needed to order the alternatives. Probabilities are related to the amount of information a possibility carries: for example, a logically true, or a deterministic outcome carries less information than a probable one. Infants can track probabilities from at least six months of age and can use this ability in order to distinguish between events Kayhan, Gredebäck, and Lindskog (2018). Infants seem to anticipate probable outcomes, but not deterministic ones, suggesting that they are aware of the tradeoff between information gain and anticipation efforts (Téglás & Bonatti, 2016). These results indicate that infants can make estimations about the information a particular event carries, in a way that resembles how information is treated in MTC (see section 1.2.3). Moreover, this capacity does not seem to depend on experience, because infants are also able to reason about one-shot, single-case probabilities (Téglás et al., 2007; Cesana-Arlotti, Téglás, & Bonatti, 2012)

However, the estimation of the informativeness of a statement, situation, or action, is a necessary but not sufficient step to organize a set. For the building blocks of exhaustification to be available, one needs a last but fundamental ingredient: ordering through entailment. Entailment has been shown to play a major role in the understanding of logical words. Chierchia et al. (2001) and Panizza, Notley, Thornton, and Crain (2013), for example, showed that 4 and 5-year-old children were sensitive to the entailment context to calculate the meaning of 'or' or of numerals; and Feiman, Hartshorne, and Barner (2019) proposed that preschool children make use of their knowledge of entailment to help them progress in the mapping of numeral words to quantities.

Despite the significance of these results and the prominent role entailment assumes in language and number theories, sensitivity to entailment per se has never been directly tested in children or preverbal infants. Such a dearth of evidence is problematic for a theory that assigns to logic a prominent role in the organization of thought.

1.4 Concluding remarks

We now want to take stock and tie the threads briefly developed in the above discussions. With these last lines about entailment we have completed our revision of what we consider to be the necessary building blocks for exhaustification, and their links to other cognitive abilities. We have illustrated infants' preverbal knowledge of logical operators (Cesana-Arlotti et al., 2018), their ability to generate sets (Feigenson & Halberda, 2004), their ability to estimate the quantity of information (Téglás & Bonatti, 2016), and the connections between the conception of alternatives and infants' causal and inferential reasoning abilities (Stahl & Feigenson, 2015). Considering the overall picture that our overview seems to paint, it looks that we are left with two main gaps in the theory we are exploring: the lack of a cohesive developmental account about words with logical valence (mainly numerals and connectives) and the absence of conclusive results about entailment abilities in children. These are the topics we will be exploring in this dissertation.

Chapter 2

The similitudes and differences in the acquisition of logical words: the case of numerals and 'and' [1]

2.1 Introduction

All throughout history, psychologists, philosophers and linguistic have debated the relationship between language and thought, but the nature and extent of the conceptual resources they share is still controversial. The spectrum ranges from those who conceive language as the source of some (Carey, 2009) or all (Van Orman, 1960) concepts to those who deny any role of language in the foundation of thought (Frege, 1918).

Central to this debate is a place where logic and language clearly meet: words with logical valence -i.e. quantifiers, connectives and numerals. There is a consensus that these linguistic expressions relate to logical functors ('and' with $\wedge$, 'all' with $\forall$, and so on). At the same time, the meaning of such words is not restricted to those of their logical counterparts. It is often assumed that these expressions have a basic, *at least* meaning, and get their other meanings essentially by pragmatic enrichment, induced by the assumption of hidden premises or contextual factors (Horn, 2006). Undoubtedly, contextual information is fundamental to our

[1] **The contents of this chapter are soon to be submmited as an article entitled: 'The similitudes and differences in the acquisition of logical words: the case of numerals and 'and'' authored by Irene Canudas-Grabolosa, Elena Pagliarini and Luca Bonatti.**

processes of mutual understanding. For example, a sentence such as (22) can be interpreted as (23) which we will refer to as *at least* meaning or as (24) which is usually referred to as 'enriched' meaning:

(22) Mary has two dogs.

(23) Mary has*at least* two dogs; indeed, she has three.

(24) Mary has exactly two dogs, and it is not the case that she has more than two.

However, the link among these meanings gets stronger when we look at them from the vantage point of some linguistic theories. Notably, Chierchia (2013) and Fox (2007) proposed that words with logical valence, such as *some* or *or*, obtain their basic, at least meaning (*at least one*, *at least one of a or b*) from the logical operator they express, but that also their enriched meaning (*some, but not all*, *a or b but not both*) derives from another, equally logical, operation. This type of inferences are sometimes called scalar implicatures (Gazdar, 1980; Sauerland, 2004; Fox, 2007; among many others). The derivation of these inferences is still debated: according to some accounts these inferences are a post-grammatical pragmatic phenomenon (Grice, 1975); according to other accounts these inferences are a grammatical phenomenon (Fox, 2007; Chierchia, 2013). Here, we will assume the grammatical account of scalar implicatures according to which the enriched meaning arise due to the application of a covert exhaustivity operator, exh, akin to 'only'. This exhaustivity operator strengthens a given sentence as much as possible; it considers the sentence and a set of relevant alternatives that differ in the quantity of information they convey and negates a subset of these alternatives. Within this framework, it thus seems reasonable to consider the meanings of words with logical valence as stemming from the same system, and thus, that it has the same origin. We will call this hypothesis the *Logicality of Language* Hypothesis.

Here, we focus on the acquisition of some of these words with logical valence and their *at least* and enriched meanings: the connective 'and' and the numerals. This choice is driven by two main reasons, which uniquely relate them. First, the meanings of these expressions share a considerable degree of overlapping: in a finite domain, expressions with logical valence composed with 'and' are either equivalent to others containing a numeral or imply expressions containing a numeral ('a dog and a dog' is equivalent to 'two dogs'; 'there is a dog and a cat' implies 'There are two animals', although the converse is not necessarily true). Second, as noticed already by Horn (1989), numerals can carry an *at least* and an enriched meaning, although theories diverge as to the exact relationship between the two (Chierchia et al., 2001; Papafragou & Musolino, 2003; Hurewitz et al., 2006; Huang & Snedeker, 2009; Barner & Bachrach, 2010; Huang et al., 2013; Feiman et al., 2019). However, also ‘and’ can carry both meanings: for example ‘and’ can receive an *at least* interpretation in ‘I saw John and Mary (but perhaps other people)”, but can also have an, enriched, ‘exact’ reading, as in ‘I saw John and Mary, and nobody else’.

Our aim is to understand to what extent expressions which share so many commonalities in structure and meaning, though belonging to two different systems, follow a similar developmental path, as it were expected if the computation of the enriched meaning of both complex conjunctions and their numerical counterparts were governed by a common primitive operation across development. We will proceed as follows. We will first review the current state of knowledge about the development and acquisition of numerals and ‘and’. We will then test specific predictions about their connections, with the hypothesis that common mechanisms underlie the construction of the meanings associated to these functors as a guide. To do this, we will test 3 year-old children with a novel paradigm adapted by the work on the acquisition of numbers. To anticipate our results, we will show that children who struggle inferring the enriched ‘exact’ meaning of numerals also have similar difficulties in understanding complex conjunctions, although the kind of errors they make seem to be different. We will finally discuss the significance of these results in the context of the Logicality of Language hypothesis.

2.1.1 The acquisition of numerals and ‘and’

Despite its apparent simplicity, the acquisition of numerals is far from trivial (see Carey, 2009). Well after having learned the words for numerals as a routine, it takes more than a full year for children to master their meanings (Wynn, 1992). They go through distinctive stages in which they seem to understand only *one* (one-knowers, around 24-30 months of age), then *two* after about 6 months, then *three*, *four*. Then, rather abruptly, they seem to generalize that to all numbers within their number list (Wynn, 1990, 1992; Le Corre & Carey, 2007). The most famous procedure that revealed this developmental progression is the ‘give-n’ task (Wynn, 1990). In this task, children who are in the middle of their numeral-mapping process are asked to hand the experimenter a growing number of objects (e.g. ‘Give me three toys’). Even those who are able to count up to *n* toys fail to provide exactly *n* toys when asked, showing the difficulty of the number-to-quantities mapping.

We recalled that this acquisition process ends when children become aware of the cardinal principle (CP), understanding that *next* in the number line means *+1* (Wynn, 1992), and thus become CP-knowers. This achievement allows them to generalize the number-to quantities mapping to all numerals. The process, which has sometimes been described as a qualitative leap (Piantadosi, Tenenbaum, & Goodman, 2012), involves the acquisition of a series of non trivial routines, and notably, the one-to-one correspondence principle, by which objects can be put in 1-1 correspondence with the numerals in the number list (Le Corre & Carey, 2007; Carey, 2009; Davidson, Eng, & Barner, 2012; Cheung, Rubenson, & Barner, 2017; Carey & Barner, 2019). Crucially, the development of such routines would depend, to a great extent, on language development (Carey, 2009; Cheung et al., 2017; Chu et al., 2020; Schneider, Sullivan, et al., 2021).

Language is indeed, deeply related to the acquisition of numerals. For example in languages with dual markers children's mastering of dual number marking correlates with their knowledge of the numeral two (Barner, Thalwitz, Wood, Yang, & Carey, 2007; Almoammer et al., 2013; Marušič, Plesničar, Razboršek, Sullivan, Barner, et al., 2016). Likewise children learning languages with no morphological plural markers (such as Chinese) are slower to become one-knowers than those learning languages possessing them and these, in turn, are slower to become two-knowers than those learning a language with dual markers (Li, Le Corre, Shui, Jia, & Carey, 2003). At the same time, language intersects with more basic aspects of numerical representation. For example, the fact that there are no languages with a 'quadral' number morphology (Marušič et al., 2016) indicates that a sharp distinction exists between how numerosities below 4 and above 4 are coded in our representational system. Indeed, number morphology and the acquisition of numerals bear strong resemblances with the object tracking system (Carey, 2009), suggesting that the representation of the numerals (1-3) can be a direct representation of individuals: thus, the representation of three dots would be identical to the representation of object A *and* object B *and* object C (Feigenson, 2011).

The potential identity between the representation of objects and that elicited by small numerals raises the possibility that numbers and the connective 'and' may be developmentally related. If a common set of semantic operations grounds the understanding of numbers and equivalent non-numeric logical expressions, then it is reasonable to expect some sort of common developmental course. A parallel argument can be made for numbers and other quantifiers such as 'some'. In this latter case the hypothesis has been explored, but, at least using the 'give-n' task, results were mixed. In some cases no evidence has been found of such a relation (Hurewitz et al., 2006; Barner, Chow, & Yang, 2009). Other findings attested a correlated development (Barner, Libenson, et al., 2009).

The link between numerals and the connective ‘and’, even though semantically motivated, has never been tested. A potential difficulty is that, at least in development, controversy still exists on what meaning for numerals children have access to, whereas there is more agreement on ‘and’. Panizza et al. (2013) showed that, similarly to adults, 3-5 year olds can access both the *at least* and the enriched interpretations, depending on their context of occurrence. On the basis of a multiplicity of methodologies, some researchers argued that children begin by assigning an exact interpretation to numerals, whereas the *at least* meaning comes as a result of an implicature (Papafragou & Musolino, 2003; Hurewitz et al., 2006; Huang & Snedeker, 2009; Huang et al., 2013). Other researchers propose that the first understanding of numerals is the *at least* interpretation, from which the enriched (‘exact’) meaning is bootstrapped (Barner & Bachrach, 2010; Feiman et al., 2019; Wagner, Chu, & Barner, 2019). These results may indicate the presence of an interpretive process governed by Gricean implicatures, as Barner and Bachrach (2010) argue, or, potentially, by some grammatical operation such as exhaustification (Chierchia, 2013).

The meaning of ‘and’ is considered less problematic, and perhaps for this reason it has been much less studied, despite the fact that conjunction is one of the main recursive devices in natural language (Lust & Mervis, 1980).

As for its acquisition, very early in development – before their third birthday – children use ‘and’ to produce complex sentences from simple ones (Bloom, Lahey, Hood, Lifter, & Fiess, 1980). Its meaning is commonly described as the equivalent to logical addition ($\wedge$), sitting at the bottom of a Horn scale (Horn, 1989) encompassing disjunction (Chierchia et al., 2001). According to Horn (1972) certain words occupy positions on lexical ordered by informational strength, e.g., <or,and>. Within this framework, ‘and’ (which lies in the strong position because it asymetrically entails ‘or’) has been assumed to take up by default only its logical meaning (Jasbi & Frank, 2017). It is also agreed that the meaning of linguistic ‘and’ extends beyond that of the logical functor, encompassing its well-known temporal flavor, and extending to a considerable range of other meanings – from an additive interpretation to temporal, causal and adversative readings.

Also some of these meanings can be subject to Gricean implicatures, or dependent on an exhaustification process, like *or*, or numerals (Beilin & Lust, 1975a). A case in point is the use of ‘and’ to link concepts and create new ones (*e.g., the dog with spots and an alien antenna*) (Bloom et al., 1980), where the same Gricean implicature or exhaustification process used in numerals to derive an *at least* and an enriched meanings can be also used to create ad-hoc concepts (Stiller et al., 2011). Even in production, ‘and’ appears closely related to its logical counterpart; other interpretations -temporal, causal, adversative- come later in development (Bloom et al., 1980). Interestingly, multiple conjunctions (‘a cat, a dog and a fish’) appear in production at around the same age as numerals (Lust & Mervis, 1980).

2.1.2 Current research question

These similitudes between numerals and ‘and’ motivate, *inter alia*, the question whether the understanding of complex conjunctions and their equivalent numerical counterparts is governed by a common primitive operation across development. Using an extension of the well-known ‘give-n’ task, we focus on three- and four-year-olds, because at this age children quickly progress in their understanding of the enriched (‘exact’) meaning of the first numerals (Almoammer et al., 2013). The issue is whether this progress is accompanied by a parallel process in understanding complex conjunctions.

In order to explore such relation, we ask whether we find a parallel development of the ‘exact’ meaning of numerals and the corresponding conjunctions: are those children who succeed in giving, say, exactly three toys when asked for three toys also able to give exactly three toys when asked to give object A, object B and object C?

Rate of success in both tasks, as well as the systematicity patterns when errors arise, will allow us to test to what extent a developmental parallel holds, as a potential sign of a common underlying semantic operation.

2.2 Methods

2.2.1 Participants

Sixty-one children (35 females, 26 males), ages ranging from 35.9 to 48.9 months (m=42.53, s.d.=3.36), from a nursery school in Vic (Barcelona) participated in the study.The protocol was approved by the Comitè Ètic d’Investigació Clínica (CEIC) del Parc de Salut MAR (2015/6094/I), Universitat Pompeu Fabra. Families were informed and gave their consent through a formulary distributed from the school.

From the participants, 8 children were excluded from data analysis (2 failed to complete the numerals and the ‘and’ task, 5 were not mother-tongue Catalan, and 1 because of experimental error). Of the remaining children (31 females, 21 males, ages 35.9-48.9 months [m= 42.7]), 48 were monolinguals Catalan, and the remaining children were bilingual Catalan-Spanish (4) or Catalan-Dutch (1). Children were given stickers for their participation. Sessions were videotaped.

2.2.2 Materials

Eight sets of ten toys (2-5 cm) were selected, belonging to the categories of cats, dogs, fish, penguins, strawberries, bananas, cars and flowers. Six of these sets were used for the experiment (three animate and three inanimate categories), one was used to elicit children’s count list, and one varied between participants. All toys were inside a carton box (40 x 20 cm) placed on a table in the middle of the room where the experiment was run. On the right side of the box (from the child’s perspective) there was a white paper with a printed red circle (20 cm). At the beginning of the session, children were given a blank DIN-A5 where they could stick the animal stickers they were given as a reward, at the beginning of the session and the end of each task.

2.2.3 Procedure

Children were tested in a small room at the school premises. At the center of the room two children-sized chairs and a table were placed. A camera which videotaped the session was positioned on one side of the table. The experimenter accompanied the child from his/her classroom to the experimental room, where she asked them to sit on one of the chairs and gave them some stickers to encourage them to participate. After that, the experiment began. Testing consisted of three tasks, always administered in the same order: the counting task, the ‘give-n’ task, and the ‘give-and’ task. After each of the two give-tasks, children received stickers as encouragement. The average duration of the session was about 30 minutes.

Counting task

This first task served to establish children’s knowledge of the counting routine and the numeral words up until ten. The experimenter showed the child one of the toy sets (not used afterwards) and, together with the child, aligned the toys on the table. She then asked the child to count the toys. Children counted twice, once starting from one side of the line, and a second time from the other side. Four participants who refused to count at the beginning of the study were administered directly the 'give-n' task and asked again to count at the end of the first task.

Give-n task

After removing the set used in the Counting task from the table, the experimenter asked the child to name all the six remaining sets of objects, so as to ensure that they knew their names. She explained that all the toys 'lived happily together' in the cartoon box, but from time to time they had to go to the red circle.

Following this introduction, she asked the child to place one toy in the red circle. After the child's action, the experimenter asked whether s/he was 'sure that there is one toy'. If the child responded affirmatively s/he was praised and encouraged. If not, the experimenter asked to 'try to fix it'. This first trial was considered a familiarization trial and was not analyzed.

Afterwards, the child was asked to place two toys in the red circle, while the experimenter closed her eyes to interfere as little as possible with the child's response. When the child was done, the experimenter opened her eyes and asked to 'count and make sure' there were two toys in the circle. If the child realized there was a mismatch, the experimenter asked to 'fix it', and the process of closing eyes-make-sure question started again. When the child was finally satisfied with the result (or stopped trying), the experimenter asked for another numeral.

The experimenter followed a titration method (Wynn, 1990; Le Corre & Carey, 2007; Barner, Libenson, et al., 2009): she asked the child to give *n+1* if they succeeded in giving *n*, or *n-1* if they failed, testing them until size 6 or up to the smallest numeral they were unable to provide in two out of three attempts. That procedure allows one to determine the highest numeral a child could give (asked at least three times) as well as the smallest numeral they could not provide. Children who reached 6 were asked 5 and 6 twice, so as to make sure their results were consistent.

The experimenter always asked children to provide generically 'toys' (see Fig. 2.1), but when they did not follow the instructions, they were either asked for specific toys instead of the generic formulation, or asked to hand the toys to the experimenter directly, instead of putting them inside the red circle.

Following Wynn (1990), children were classified as n-knowers if they could give *n* two out of three times when asked for *n*, and did not give *n* more than the 50% of the times when asked for other numerals than when asked for *n*. For example, a child who systematically gave two objects when asked for two and for three was classified as 'one-knower', because she gave 'two' as many times when asked for other numerals as she did when asked for 'two'.

Give-and task

In this novel task, whose procedure was patterned upon the 'give-n' task, children were prompted with a conjunction asking them to pick specific toys identified by their category membership (e.g. *Can you put a cat and a dog inside the circle?*). Then, as in the 'give-n' task, the experimenter asked them whether they were sure they had put 'a cat and a dog' in the circle (for a comparison of the two tasks, see Fig. 2.1). If their answer was negative they were asked to correct it; if it was affirmative they were congratulated, and a new trial, following the titration method, started.

In order to reduce the role of memory, which is more taxed than in the 'give-n' task at a parity of array size (because the objects must be individuated by category), children were repeated the instructions as many times as they asked for. If children succeeded, the task continued by adding a new conjunct to the conjunction (e.g., the experimenter asked for *a cat and a dog and a penguin* after children succeeded with *a strawberry and a flower*). Alternatively, the conjunction was simplified if children failed with the previous conjunction (thus, in the previous example, children could be asked for *a cat*). We chose to link each new conjunct with an 'and', instead of eliding it, in order to make the conjunctive structure more salient.

Children were asked up to 5 conjuncts. Notice that we always had an unused set of toys, different for each child. Thus, children could not determine what set the last object belonged to by exclusion, if they ever reached that level. The experiment stopped at the lowest number of conjuncts children failed to give for two out of three times. The classification criteria were the same as in the 'give-n' task.

You can see an example of the procedure in the Supplementary Material folder.

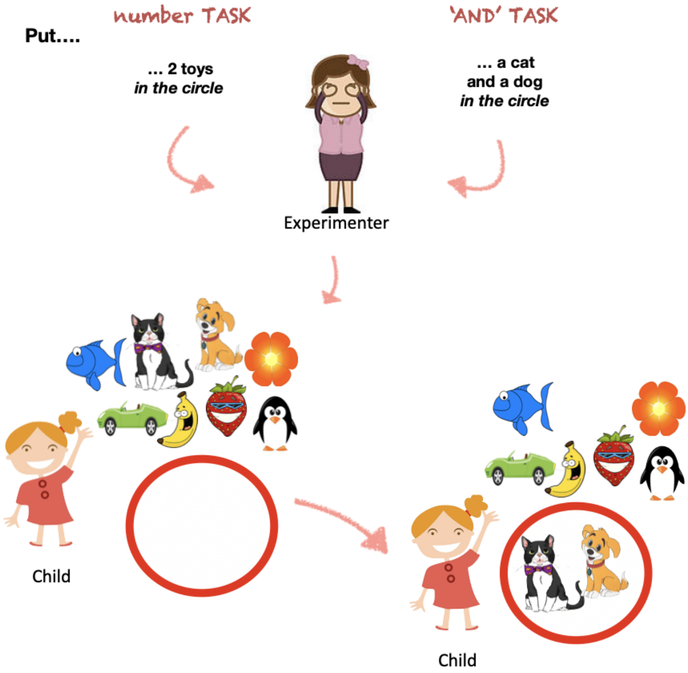


Figure 2.1: Schema of the two tasks ('give-n' task and 'give-and' task)

2.3 Results

The Counting task confirmed that all children understood the counting routine and were able to count until 6, the highest number they were asked. Thus, data from all of them were analyzed. Children's performance, however, was markedly different in the other two tasks (see Fig 2.2). Overall, children performed better in the 'give-and' than in the 'give-n' task (give-and: m= 3.17, s.d=1.28, give-n: m= 2.68 , s.d.=1.65; t(52) 2.14, Wilcoxon paired sample test, V=259, p<.038). In the 'give-n' task, children distributed unevenly across the experimental levels (chi-squared (6) = 13.25, p< .04), with most of them being n-counters for n <4; however, a sizable minority did act correctly even with n=5 and n=6, most likely because they were CP-knowers.

We checked whether the distribution of n-knowers was in accord to previous literature. To this purpose, we compared our data with n-knowers in a dataset provided by the The Language & Development Lab at UCSD, kindly provided upon request, including results from published articles (Barner, Libenson, et al., 2009; Almoammer et al., 2013) as well as ongoing studies. We restricted the analysis to *n*-knowers (CP-knowers excluded) of the same age as our group, tested with a comparable methodology. The distribution of *n*-knowers of the same age (learning English, a language with singular and plural but no dual, as Catalan) as our sample was no different from that of this database (p= .619, Fisher exact test). Thus our sample of *n*-knowers seems to be in line with what had to be expected given previous work.

In the 'give-and' task, most of the children were unable to act correctly when the conjunct sentence had more than three conjuncts (35 out of 53). Remarkably, of the 18 who were able to go beyond 3, 12 participants managed to act correctly even when the sentence contained the maximum number of conjuncts. An account appealing to memory limitations would predict no or few successes with such long conjunctions.

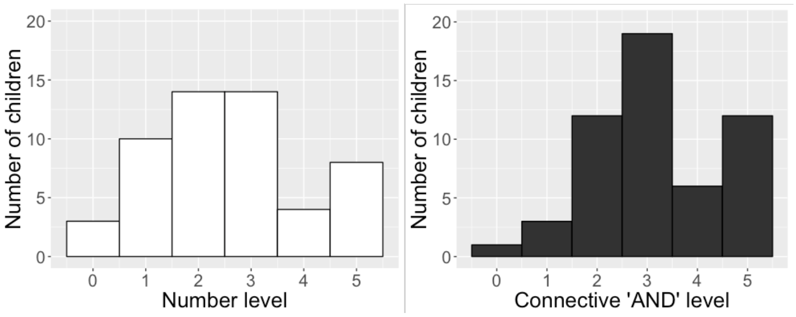


Figure 2.2: Histograms of number level (give-n task) and connective 'and' level (give-and task).

Next, we explored the relation between the two tasks. The exact understanding of numeral (the enriched meaning: 'exactly two and nothing more') correlated with the exact understanding of 'and' (enriched meaning: 'only a fish and a car and nothing more') (Spearmann cor.; rank= .4, p< .004; Fig. 2.3), even when controling for age (proportional odds logistic regression; main effect for give-and, t(52)= 2.56, p=.01; main effect for age, t(52)= 7.77, p<.001).

It is worth mentioning, however, that a considerable number of children succeeded in the 'give-and' task even with size 5, and yet failed the 'give-n' task with n higher than three (9 out of 12, or the 17% of the total participants). By contrast, those children who succeeded with high numerosities equally distributed across all the and-levels. Thus, success at high quantities in the first task was not predictive of success in the second task. Besides being intrinsically interesting, this difference in behavior in the two tasks shows that children's failures could not be attributed to a common memory limitation for larger arrays.

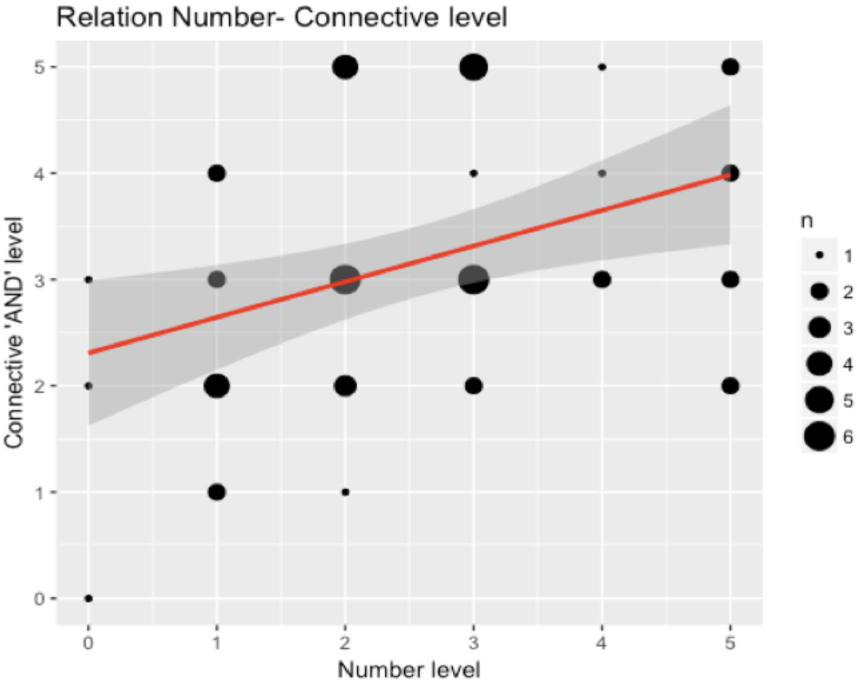


Figure 2.3: Relation between the understanding of numerals and the understanding of conjunctives with 'n' conjuncts

Finally, we analyzed errors. We classified responses according to whether children gave more, less, or the exact number of objects that they were asked to put inside the circle. Responses in the 'give-and' task, even when the exact number of objects was provided, were further classified according to whether children gave the exact type of objects required or made category mistakes (see table 2.1 for examples of the classification).

Type of response		**Give-me-n task**	**Give-me-and task**
Less	**Definition**	Giving less objects than required	
	Example	asked for 3 objects, the child gives 2	asked for 'a penguin and a fish and a banana' the child gives a banana and a fish.
More	**Definition**	Giving more objects than required	
	Example	asked for 3 objects, the child gives 4	asked for 'a penguin and a fish and a banana' the child gives a penguin, a fish, a banana and a cat.

Exact	**Definition**	Not an error, giving the number of objects they were required.	**N+C+(number and category exact):** Not an error, giving exactly the objects they were required.
			N+C- (number exact but category wrong): giving the number of objects they were required, but making a mistake on the identity of some of them
	Example	asked for 3 objects, the child gives 3	**N+C+(number and category exact):** asked for ‘a penguin and a fish and a banana’ the child gives a penguin, a fish and a banana.
			N+C- (number exact but category wrong): asked for ‘a penguin and a fish and a banana’ the child gives a penguin, a fish and a car.

Table 2.1: Classification of children’s responses. Each trial of each child was classified in one of these categories.

We classified participants who did not commit errors as ‘exact’ in the ‘give-n’ task and ‘N+C+’ in the ‘give-and’ task. Interestingly, and somewhat surprisingly, we found very few category errors in the ‘give-and’ task (N+C-, giving the exact number, but the wrong category). Again, this result seems to indicate that the pattern of errors we document was not due to memory task demands alone. Because we found no effect of the limited N+C- errors, we collapsed the two categories in successive analyses.

Focusing on errors, we selected the trials with errors, and classified participants as error-systematic if they made the same error at least 2/3 of the times. There was considerable systematicity in the type of errors participants made. Fifty-two children were error-systematic in the ‘give-n’ task and 42 in the ‘give-and’ task (Table 2.2). Most participants gave more objects than required, especially in the ‘give-n’ task. Although the titration method may contribute to this error, its presence is also a sign that children at this age are more likely to adopt an *at least* reading of the expressions involved and have difficulty in switching from it to the enriched meaning, an operation which may require exhaustification.

Systemacity pattern		**'Give-n' task**	**'Give-and' task**
less		14	13
more		33	21
exact	**N+C+**	5	6
	N+C-		2
Unclassified		1	11
Total		**53**	**53**

Table 2.2: Number of children that were classified in each of the categories. Children were classified as belonging to one category if, when they made an error, they made the same one at least 2/3 of the times. Please notice that children who fell into the 'exact' category in 'give-n' task and the N+C+ in the 'give-and' task were those who did not make mistakes on the task. Children labeled as 'unclassified' where those who were not consistent on their errors.

Particularly important for our investigation was the question whether systematicity for one task would predict systematicity for the other. This was not the case: error systematicity had little or no positive correlation across tasks (Goodman-Kruskal τ = 0.02), and the negligible relation was in the opposite direction. This was true regardless of whether we analyzed only error-systematic children or we included unclassified children, or even exact children (see table 2.3). We comment on the meaning of this result below.

And **Numb.**	**Less**	**More**	**N+C+**	**N+C-**	**Uncl.**	**Total**
Less	2	7	2	0	3	**14**
More	9	14	3	1	6	**33**
Exact	2	0	1	1	1	**5**
Unclassified	0	0	0	0	1	**1**
Total	**13**	**21**	**6**	**2**	**11**	**53**

Table 2.3: Number of children who fall into each Systematicity patterns for the 'give-n' and 'give-and' tasks. Identical systematicity patterns would imply that most data to fall into the diagonal of the table (highlighted). However, they do not.

2.4 Discussion

The relation between language and logic has always been a crucial topic in philosophy, linguistics and psychology. A way to address it is to closely look at natural language words which have logical valence, and to understand their interconnections in development. In this paper, we started by looking at the relationship between words that in many contexts have equivalent meanings, in order to find cues on potential common logical processes underlying the acquisition of their linguistic meaning. We were inspired by the hypothesis that several inferential behaviors generally attributed to pragmatic factors may in their nature be fruit of grammatical operations over logical forms, as the Logicality of Language hypothesis holds.

We introduced a novel task, the ‘give-and’ task, patterned upon the ‘give-n’ task, and we compared children’s mastering of numerals to their understanding of equivalent expressions with multiple conjuncts. The results revealed the existence of a relation between the understanding of numeral words and the connective ‘and’: children who were struggling to infer the enriched ‘exact’ meaning of numerals were also struggling to infer the ‘only A and B and nothing else’ meaning of conjunctions with a number of conjuncts corresponding to the numerals they had difficulty with.

At the same time, the fact that the performance in the ‘give-and’ task was higher than in the ‘give-n’ task, and that a sizable amount of children had no difficulty in understanding long complex conjunctions and yet were still struggling with the corresponding numerals, exclude that the whole pattern of results can be explained by simple memory task demands. We also found that in both tasks children seemed to systematically make more errors compatible with an *at least* interpretation, suggesting that they struggled to come up with the enriched interpretation of these expressions. Thus, the correlation between the understanding of numbers and of complex conjuncts, as well as the number of errors compatible with an *at least* interpretation, seem to point at a common origin in the acquisition of the first meaning children assign to both numerals and multiple conjunctions.

These data may be evidence that a common operation underlies the understanding of these expressions. Specifically, exhaustification, which rules the passage between the one and the other meanings of a logical expression, can be at the source of children’s patterns of errors and successes. The acquisition of the one-to-one principle, fundamental to understand the numerical series, is not sufficient to explain why and in which contexts a number is interpreted in an *at least* or enriched meaning, nor is it sufficient to explain children’s understanding of complex conjunctions. Exhaustification can predict some of the conditions in which these changes of interpretation occur; children's hesitations between *at least* and enriched interpretations of expressions belonging to different domains may be caused by the same difficulty: the acquisition of this operation, whose mastery affects a wide range of linguistic expressions, including numbers and 'and'.

Different explanations have been elaborated in the literature in order to account for the difference between adults and children in interpreting the expressions under study. A particularly relevant proposal for us is the so-called Alternative-base approach, which our proposal can be considered as complementary to. According to the Alternative-base approach, the source of difficulties for children lies in the retrieval of the alternatives that are required to compute the inference fixing their exact meaning (Barner et al., 2011; Tieu et al., 2016; Skordos & Papafragou, 2016). This proposal holds that children may fail in accessing the alternatives, and that such failures are scale-specific, rather than due to pragmatic immaturity or to the inability to associate items belonging to the same scale. This proposal found some support by results showing that when the alternatives are made easily accessible, either by the linguistic context (see for instance Tieu et al., 2016; Pagliarini, Bill, et al., 2018; Barner, Hochstein, Rubenson, and Bale, 2018; Gotzner, Barner, and Crain, 2020), or by the extra-linguistic context (Barner et al., 2011) or else by the experimental settings (Skordos & Papafragou, 2016; Foppolo et al., 2012), children derive scalar inferences more readily.

Our proposal is not alternative to the Alternative-base approach, but it examines this process from a different perspective. While most of the previous accounts focus on which alternatives children can consider (surely a fundamental factor), here we considered another step of the process: exhaustification. Because exhaustification is a grammaticalized operator that negates the alternative computed from an original proposition (e.g. Spector, 2007; Fox, 2007), it is plausible to think that two steps are required in order to successfully exhaustify an assertion: first the alternatives must be accessed, and then exhaustification has to be applied on them in order to force the enriched meaning out of the *at least* meaning of the original assertion.

If such primitive operation exists, then once children have access to exhaustification and know how to apply it, they will apply it across logically equivalent expressions. This is what the common difficulties children had with conjunctive and equivalently complex numerical expressions suggests.

This being said, we would like to stress that by no means does our hypothesis imply that there are no specificities in the development and understanding of the numerical system. In fact, our data suggest it as well: the parallel between numbers and conjunctions seems to break down when numbers, expressing set sizes beyond the subitizing range, require the understanding of the counting principle. ‘and’ too has its specificities, in particular because it can link together elements of different categories by simply listing them – something that numbers cannot do without climbing up the category ladder (‘A cat and a dog’, *‘Two animals’*).

Indeed, we documented that although errors were systematic within each task, systematicity in the one task does not predict systematicity in the other task. This, we submit, is due to the specificity of the two domains. That they are different is clear also from other developmental evidence. For example, the well-established incremental process of acquisition of the numeral sequence, including the explosion in the understanding of the sequence once the CP principle is acquired, has no correspondence for ‘and’: as it were, a ‘CP-and principle’ does not exist. Thus, our research contributes to understand which operations may be common sources of the origin of meaning of these linguistic expressions, without obliterating the differences between domains.

This explanation predicts that ‘and’ systematicity patterns should be reasonably constant regardless of the CP-knowing status of children, whereas systematicity patterns in numerals will change when the CP-status is attained. Due to the small number of CP-knowers in our sample we could not test this hypothesis, which is a topic for further research.

The hypothesis we propose generates further predictions. In particular, it predicts that the specific exhaustification operation required by the task will determine the type of response and, potentially, the type of errors that children make. Exhaustification is a function of the syntactic context in which the expression to be interpreted appears. In our task (‘Give me x’) the operation should force an enriched interpretation. As suggested, a reason of the difficulty children encounter in fixing the meaning of numerical and conjunctive expressions can depend on their incomplete mastery of this operation. However, in contexts where exhaustification is not mandatory, such as in Downward Entailing Contexts (e.g., ‘If you give me three apples I am happy’), children’s intuitive interpretation corresponds to the meaning predicted by the grammatical account of scalar implicatures (at least three apples) (Chierchia, 2013), which is also the meaning adults attribute to them (Panizza, Chierchia, & Clifton Jr, 2009). Further research will explore this hypothesis.

In conclusion, our study opens the possibility that the acquisition profile of ‘and’ and numerals stems from the acquisition of a common principle, operating on top of the idiosyncrasies of the domains. Under this perspective, an important developmental question is to understand the primitives that make this and other logical/semantic operations available to children, to describe how are acquired and to map what consequences in children's reasoning with expressions with logical valence they engender.

Chapter 3

Scalar Implicatures in adults: the case of numerals and connectives [1]

3.1 Introduction

One of the main functions of language is to convey meaning. However, the way meaning is encoded and transmitted is a very complex phenomenon. Acquiring the lexical content of words is an obviously necessary, but not sufficient step: words alone do not convey thoughts. The minimal units that encode truth-bearing structures are syntactically well-formed combination of words, with the inferences they engender. In this multi-facet process, some theories assign a particularly important role to the logical scaffolding of language (e.g.Lakoff, 1969; Beilin and Lust, 1975a; Chierchia and McConnell-Ginet, 2000). According to these views, patterns of implication of logical nature may constrain the meaning of linguistic expressions in conversation. Thus for this approach, understanding and communicating meaning requires an interplay between language and logical operations within language. We will call these the Logicality of Language (LOL) theories. A clear place where logic and language meet is in linguistic connectives and quantifiers. There is little doubt that linguistic expressions, such as 'and', 'or', 'if', or 'All', have a

[1] **The contents of this chapter are soon to be submmited as an article entitled: 'Scalar Implicatures in adults: the case of numerals and connectives' authored by Irene Canudas-Grabolosa, Elena Pagliarini, Gennaro Chierchia and Luca Bonatti.**

relation with the logical functors $\wedge$, $\vee$, $\rightarrow$ and $\forall$. At the same time, it is well known that sharp differences exist between the meaning of these functors and the use and meaning of their closest counterparts in natural language. Such differences have been often taken as a sign that logic has little role in our understanding of these expressions, whereas mostly conversational and pragmatic factors constrain their interpretation. However, these functors seem to possess a considerable degree of systematicity. Quite interestingly, both the natural language words expressing quantifiers and those expressing connectives possess a readingwhich more closely corresponds to their logical counterparts, although they can also acquire a more enriched meaning. Thus, when we hear a sentence like:

(25) Mary has a cat or a dog.

'at least' two interpretations are possible. In one interpretation, the speaker intends to convey that Mary has 'at least' one of the animals and possibly both the cat and the dog (akin to the logical meaning of $\vee$, true if one of the disjuncts is true). In another interpretation, the speaker means to convey the fact that Mary has only has one of them, and for example s/he does not know which one (or a more enriched meaning, true if one of the disjuncts is true but not both). This duplicity is not a case of accidental homonymy; in fact, it exists in all studied languages (Singh et al., 2013)), a fact hardly explainable as the result of the miraculous convergence of arbitrary lexical mappings. For this reason, among others, LOL and other theories assume that logical functors may have a basic, 'at least' meaning close to the one described by logic. From this one, the enriched meaning is derived by means of a scalar implicature, a process in which a scale of alternative interpretations is created, organized in terms of informativeness, and then prunned by negating the more informative terms in it to converge to the relevant meaning (Noveck, 2001; Chierchia et al., 2001; Chierchia, 2013). While this process is influenced by pragmatic (Pouscoulous et al., 2007; Guasti et al., 2005) and extralinguistic Fairchild and Papafragou (2018) factors in the context of the situation in which an utterance is expressed, its systematicity across languages may suggest that also structural aspects of the sentences play an important role. This is the line we will pursue here, specially focussing on the role of logical context to explain the distribution of meanings. We will base our research on the theory of exhaustification (Chierchia, 2013), a theory that stresses the importance of semantic rather than pragmatic factors in scalar implicatures, based on the observation that the process needed to get access to the enriched meaning can also explain purely syntactic phenomena, like the distribution of negative polarity items.

For the enrichment process to work, two interdependent factors are needed: the generation of a scale, and the negation of the alternatives belonging to that scale. A scale is needed because without it, the space of alternatives is not defined. For example, expressions such as quantifiers, modals, or numerals, are all part of different Horn scales (Horn, 1989), sets of alternative expressions from the same grammatical category ordered in terms of their semantic informativeness. A Horn Scale provides the alternative terms needed to frame a scalar implicature. Then, given a scale, one needs to exclude the irrelevant alternatives, and without the restriction of the set of alternatives, no enriched interpretation would ever arise. Thus this operation, called in our framework of reference *exhaustification*, is also required. We take these factors in turn.

3.1.1 Theories of Scalar Implicatures

Scalar implicatures have been studied since Grice (1975). Their traditional interpretation is linked to his pragmatic framework, and in particular, to the Cooperative Principle of quantity ('make your contribution as informative as required, but not more informative than required'). In this interpretation, if a speaker chooses to produce a less informative term it is because she has no evidence that the stronger statement holds; in our example (25), the speaker uses 'or' because she does not know whether Mary has both a dog and a cat; and in consequence she could not use the alternative, stronger sentence 'Mary has a dog *and* a cat' (Noveck, 2001). An alternative, also pragmatic, hypothesis denies that scalar words involve logical inferences, and rather favors the idea that they are ambiguous between the two meanings (Ariel, 2015), potentially determined by relevance (Bott & Noveck, 2004; Breheny et al., 2006; Dieussaert, Verkerk, Gillard, & Schaeken, 2011). According to this view, all conversational exchanges are based on the principle of trying to gain the maximum effect with the minimum effort, and scalar implicatures occur only when the effort to draw them is justified by the situation. Thus, 'Mary has a cat or a dog' would mean 'Mary has one of the two animals and not both' only when the speaker expects the hearer to draw an interpretation that makes this meaning relevant, that is, when there is an implicit/explicit question as to whether Mary does own both animals.

Because scalar Implicatures are governed by pragmatic principles, according to these approaches the inference has to be calculated globally, over the full speech act. This globality constraint implies that implicatures cannot be embedded in more complex sentences. That would mean that for a sentence like:

(26) Every child has a cat or a dog.

Would have, as alternative:

(27) Every child has a cat and a dog.

The Gricean reasoning would lead to the conclusion that the author of (26) does not believe (27)to be true -otherwise s/he would have said so. Therefore, the meaning of (26) would be (28):

(28) Every child has a cat or a dog and not every child has a cat and a dog.

However, Chemla and Spector (2011) 's work show that another meaning is available for speakers:

(29) Every child has a cat or a dog but not both.

The meaning conveyed in sentence (29) can only be derived if the alternative considered is not the entire sentence (27) but the clause 'a cat and a dog'. In other words, that scalar implicatures can be embedded.

For this reason it has been proposed that scalar implicatures are governed, albeit not exclusively, by semantic and syntactic factors, rather than being pragmatic (Fox, 2007; Chierchia, 2013). In this latter interpretation, grammar provides a covert operator, a sort of silent 'only', that would associate with scalar terms and would be responsible to trigger the enriched reading. With such an operator, scalar implicatures can appear and be interpreted in embedded contexts just as in simple phrases. Central to this approach is the observation that the distribution of meanings of an expression depends on the upward or downward nature of the entailment context it licenses. Entailment is a relationship between two sentences such that if the first is true, the second will necessarily be true (Chierchia & McConnell-Ginet, 2000). For instance, sentence (30) would entail sentence (31):

(30) If you open the door, I'll take a dog outside.

(31) If you open the door, I'll take an animal outside.

That is because we can infer (31) from (30), since in all situations in which (30) is true, so will be (31. In monotonic contexts, we can distinguish between two entailment contexts. Sentence (30) is an example of an upward entailing (UE) context, since it licenses inferences from the set (dog) to the superset (animal). The consequent of conditional sentences, as in the case of (30), is an upward entailing context. On the contrary, inferences can also be licensed from the set (dog) to the subset (Dalmatian):

(32) If you take a dog inside, I'll close the door.

(33) If you take a Dalmatian inside, I'll close the door.

As one can easily see, sentence entails sentence , since in all situations in which I commited to close the door if you take a dog outside, you will expect me to do so if you take a specific dog, in that case a Dalmatian. These contexts, such as the antecedent of the conditional, that license inferences from the set to the subset are called downward entailing (DE) contexts.

This division of contexts according to the entailment pattern is relevant to account for several linguistic phenomena. For example, the distribution of Negative polarity items, like *any* and *ever* can be explained by noticing that such quantifiers are only permitted in DE contexts (Chierchia, 2013). For scalar implicatures, the enriched meaning (a cat or a dog but not both) tends to be easier to derive in UE contexts (example (34)):

(34) If you open the door, I'll take a cat or a dog outside.

The 'at least' meaning (a cat or a dog, or both), on the other hand, is more commonly licensed in DE environments(Ladusaw, 1979; Chierchia et al., 2001; Gualmini, Meroni, & Crain, 2003; Chierchia, 2013, 2004; Panizza, Chierchia, & Clifton Jr, 2009):

(35) If you take a cat or a dog outside, I'll close the door.

Multiple experimental investigations have attempted to disentangle between these approaches. Some studies suggest that while both meanings are accessible, the restrictive meaning requires an extra step, carrying some processing cost. This could be attributed to the derivation of the implicature (**dieussaertal11; bottbaileygrodner12; tomlinsonbaileybott13**; Bott and Noveck (2004), Breheny et al. (2006), Huang and Snedeker (2009); but see also **grodneral10; brehenyfergusonkatsos13** for opposite views). Crucial to our work, a growing amount of literature indicates the importance of grammatical factors in the interpretation of the quantifiers (Chemla & Spector, 2011; Hartshorne, Snedeker, Liem Azar, & Kim, 2015) , numerals (Panizza, Chierchia, & Clifton Jr, 2009; Panizza et al., 2013) and 'or' (Chierchia et al., 2001; Pagliarini, Bill, et al., 2018). Panizza, Chierchia, and Clifton Jr (2009) studied adults' interpretation of numerals in UE and DE contexts. A DE context was achieved by embedding the numeral in the antecedent of a conditional, or in the scope of a universal quantifier. Both with offline and online measures, for both universal and conditional sentences they found a higher acceptance of an enriched (exact) reading of numeral in UE than in the DE contexts. Also, reading time for the numerals was longer in an UE context, indicating that processing resources were required when the context triggered an enriched interpretation. A similar effect was found in 3 to 5 year old children (Panizza et al., 2013).

Also the interpretation of 'or' has been found to be sensitive to the context in which it is embedded. Chemla and Spector (2011) showed that when a scalar item (either 'some' or 'or') occurred in the scope of a universal quantifier, in a sentence-to-picture matching verification task the picture corresponding to the enriched meaning was more accepted than that corresponding to the 'at least' meaning if the scalar occurred within the scope of the universal quantifier (which creates a UE context), whereas that this tendency reversed in DE contexts. A similar effect has also been found in 5 year-old children (Chierchia et al., 2001; Singh et al., 2013), although in these studies some inconsistencies in the responses of the control adults surfaced, as they interpreted 'or' in its logical sense even when the entailment context favored the enriched meaning (e.g. Singh et al., 2013) Overall, these results suggest that speakers keep track of the amount of information that sentences carry and the asymmetrical entailments they imply (Hartshorne et al., 2015). On this background, the LOL hypothesis sees scalar implicatures as procedures to maximize the information carried by sentences (Chierchia, 2013). This process is sensitive to the upward/downward context in which a logical expression appears. Thus in an UE context, an enriched meaning of 'or' is more informative than the logical one, as it allows us to rule out an extra possibility. Let's see why in example (34), repeated here as (36):

(36) If you open the door, I'll take a cat or a dog outside.
  a. Possibility 1: If you open the door, I'll take a cat outside.
  b. Possibility 2: If you open the door, I'll take a dog outside.
  c. Possibility 3: If you open the door, I'll take a cat and a dog outside.

In example (36) it can be seen that ruling out possibility 3 makes the mening of (36) more restrictive (i.e.- it makes the sentence (36) logically true in less situations) than not ruling it. Thus, enriched 'or' is more informative, in that UE context, than ‘at least’ ‘or’. However, if the same functor is embedded in a DE context such as in the antecedent of a conditional that would not be the case:

(37) If you take a cat or a dog inside, I’ll close the door.

a. Possibility 1:If you take a cat inside, I’ll close the door.

b. Possibility 2:If you take a dog inside, I’ll close the door.

c. Possibility 3:If you take a cat and a dog inside, I’ll close the door.

Interestingly, in example (37) ruling out possibility 3 does not make the meaning of (37) more restrictive (in fact, it makes the sentence (37) logically true in *more* cases, so it turns it into a *less* restrictive meaning). Thus enriched ‘or’ is *less* informative, in that DE context, than ‘at least’ ‘or’. Crucial to this interpretation , informativeness is determined through the construction of a space of possibilities, by computing the truth values of a sentence (similar to what the Mathematical Theory of Communication (MTC) claims. See section 1.2.3). Such approach heavily relies on the logical meanings of the words, and would be able to explain scalar implicatures relying on the compositional semantics of the sentence, without appealing to pragmatical factors. Yet, a systematic study of how adults’ interpretation of scalar items is modulated by the entailment context is lacking. Our aim is to further test the LOL hypothesis by filling this gap.

3.1.2 Aim of the present study

We test how Upward and Downward Entailing contexts affect the interpretation of conjunction, disjunction, and numerals embedded in conditionals. Complex conditionals are a fundamental test case. First, they provide a well controlled test to examine how the interpretation of scalar expression is affected by compositional contexts. Then, they offer a way to tests specific predictions of the LOL hypothesis. Their antecedent creates a DE, while their consequent an UE, context. By switching the position of the embedded phrase between the antecedent and the consequence of a conditional, one can control sentence complexity while studying the specific effect of upward and downward contexts in the interpretation of the sentence. We test both members of the Horn scale <or, and> in order to ascertain whether potential changes of interpretation due to embedding were specific to the weakest logical functor or could be detected across the whole scale. Sensitivity to the entailment context for ‘or’ has been studied in children (Chierchia et al., 2001) but, to our knowledge, no systematic study exists for adults. ‘And’ has never been tested, whether in adults or children. We compare logical functors with numerals because they provide a useful point of comparison, as they are sensitive to the entailment context (Panizza, Chierchia, & Clifton Jr, 2009); and in many contexts are equivalent to 'and', thus offering a direct test of potential differences in the interpretation of equivalent expressions in polarity sensitive contexts. In order to make the comparison of numerical and logical expressions closer, we target both simple conjuncts/disjuncts ('a cat and/or a dog') and chains of 'or' and 'and' ('a cat, a dog and/or a fish') in the embedded expression. In this way, we can create a parallel between logical expressions (e.g. A and B) and numerals (two), which also serve as a visual control, since it allows us to evaluate the influence of the logical expression in the response of participants to the exact same visual scene. LOL predicts that participants should tend to favor the inclusive (logical) meaning of 'or' ('A or B or both') and the lower-bounded meaning of 'numerals' (e.g., "at least' 2, possibly more') when these functors are embedded in in DE, when compared to UE, contexts, The case of 'and' is interesting. As it is at the lower end of a Horn scale <and, or> (Horn, 1989), its interpretation should not change according to the embedding context. However, as it bears very close semantic similarity with the corresponding numerals, it is possible that, like them, it switches from an exact interpretation ('A and B and nothing else') to an 'at least' interpretation ('A and B and maybe something else'). This operator, which has not been studied in these contexts to our knowledge, will inform us how strongly the DE/UE contexts can twist the semantic arm of the elements of scales. We test these predictions in Experiment 1. Experiment 2 addresses the role of (other) pragmatic phenomena that may selectively affect participant's responses in Experiment 1, which we call the *partial satisfaction* and the *magnitude* effects. We finally discuss the implications of the results for theories on Scalar Implicatures and logical cognition in general.

3.2 Experiment 1

3.2.1 Methods

Participants

Thirty-two Catalan-speaking adults were recruited at the Universitat Pompeu Fabra (Barcelona). They were from Barcelona and its surroundings. Two of them were excluded from the analysis because they made more than the 20% erroneous answers to the filler items. For the remaining 30 participants (22 females), ages ranged from 18 to 24 years (M= 21). Participants received €10 for their participation.

Procedure

We created a ‘scaled acceptability’ judgment task. In this task, participants were asked to grade using a scale from 1 to 5 the actions that two characters performed in a series of stories, based on what the characters of the story had previously announced. Participants had to read the characters’ utterances, watch them perform the action and finally grade such action appropriately. The experiment began with a presentation of the two characters, the scale of acceptability, and an explanation of their task. Then they were given 5 familiarization trials, designed to tune their responses to the 5-level scale. Afterwards, the proper test began. The task was administered using PsyScope X B88 (Cohen, MacWhinney, Flatt, & Provost, 1993). Participants were tested in a quiet room in the Center for Brain and Cognition Laboratory (https://www.upf.edu/web/cbc), at Universitat Pompeu Fabra (Barcelona). They wore headphones for the entire experimental session. The study lasted 40 minutes.

Materials

The experiment included 48 stories (16 per each logical operator ‘and’, ‘or’, ‘numeral’) and 26 fillers. Half of the fillers contained a conditional sentence but deprived of the logical operators under study (e.g. if it rains we take an umbrella). The other half contained a sentence with one or two of the logical operators under study, but no conditional (e.g. I take two apples and she takes a banana). Thus, participants saw a total of 74 stories, presented in pseudorandom order. Each filler was built upon a different narrative, whereas the 16 stories per logical operator were variations of six different narratives, randomized across conditions. All stories followed the same pattern. They started with a bar that filled as countdown and an attractor sound, and then the story began showing a girl and a boy. In the test trials, one of the two characters attracted participants’ attention to a set of objects (toys or food). Then, the second character specified that his/her goal was to do something with them (take them out, put them in the fridge…). Finally, the first character uttered the conditional target sentence, which contained one of the target logical operators (’and’, ‘or’, ‘numeral’). The sentence was a request to perform an action on some of the objects (from 2 to 4). The logical operator could be in the consequent (UE contexts) or in the antecedent of the conditional (DE context).

Each operator appeared in UE contexts in eight stories and in other eight stories in DE contexts (see Table 3.1).

Logical operator	UE context	DE context
Numerals	If you open the door, I'll take 2 toys out.	If you take 2 toys in, I'll close the door.
Or	If you open the door, I'll take the cat or the dog out.	If you take the cat or the dog in, I'll close the door.
And	If you open the door, I'll take the cat and the dog out.	If you take the cat and the dog in, I'll close the door.

Table 3.1: Target sentences per each logical operator (numerals, 'or' and 'and') and context (upward entailing context and downward entailing context)

After the target sentence, the two characters performed the action described by the conditional. The character uttering the phrase without the target logical operator ('non-actor') always performed the exact action stated by the sentence (i.e. If his/her part was to 'close the door', s/he would always do it, irrespective to what the other character had done or would do). For the character acting on the logical operator phrase ('actor'), there were 3 different conditions:

- 'Exact' condition: the actor handled the exact number of objects mentioned in the target sentence (e.g., manipulating 2 objects when the target sentence mentioned 2). Theoretically, that condition would correspond to the 'at least' meaning of 'or' and to the enriched meaning of numerals and 'and'.
- 'More' condition: the actor handled one object more than required (e.g., using 3 objects when 2 had been mentioned). Theoretically, this would correspond to the 'at least' interpretation of numerals and 'and', and of 'or'.
- 'Less' condition: for conjunction and 'numeral', the actor handled one object less than required (using 2 objects instead of the 3 mentioned). That would correspond, for numbers and 'and', to an error, since it fails to comply with the meaning of the logical operator. For disjunction, the actor manipulated only one of the requested objects, which would correspond to the enriched (exclusive) interpretation.

Table 3.2 shows a summary of the correspondence of our conditions to the 'at least' and enriched meanings of the logical operators.

	Less	**Exact**	**More**
Numerals	Error	Enriched	At least
Or	Enriched	At least	At least
And	Error	Enriched	At least

Table 3.2: Comparison between our conditions (less, exact and more) and the enriched and 'at least' meanings of the logical operators used in the study ('or', 'and' and numerals)

Context (UE and DE), Logical operator (or, and, numeral) and Condition (Exact, More, Less) were fully crossed, thus yielding a 3x3x2 experimental design. The resulting stories were repeated twice. In addition, four sentences per logical operator (2 per context) implemented a control condition. In it, the conditional was not satisfied, either because the actor did not provide any object or because the non-actor made an action that did not comply with the meaning of his/her other part of the sentence (be it antecedent or consequent, depending on the context). For example, if the target sentence was 'If you open the door, I'll take the cat and the dog out' or the inverse 'If you take the cat and the dog in, I'll close the door', the non-actor would not open/close the door but the window. Participants were asked to grade the actor's actions (relative to the target phrase in the sentence) on a Likert-like scale from 1 to 5 (see fig. 3.1), by pressing the corresponding key on the computer keyboard. Participants could also press an 'I don't know' button. This ensured that the middle values of the scale were not used to express a state of ignorance, which could occur either because they had forgotten the sentence or because they had missed any part of the story. Fig. 3.2 shows an example of a story:

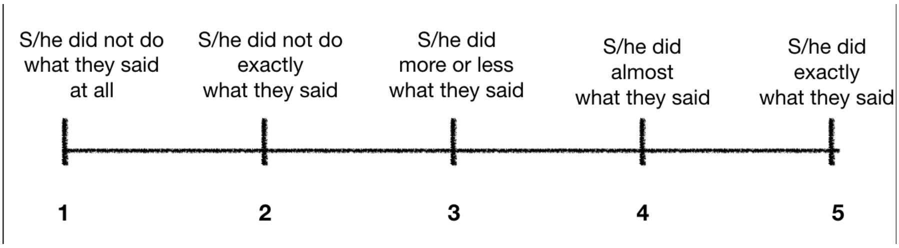

S/he did not do what they said at all
S/he did not do exactly what they said
S/he did more or less what they said
S/he did almost what they said
S/he did exactly what they said
1
2
3
4
5

Figure 3.1: Scale of response.

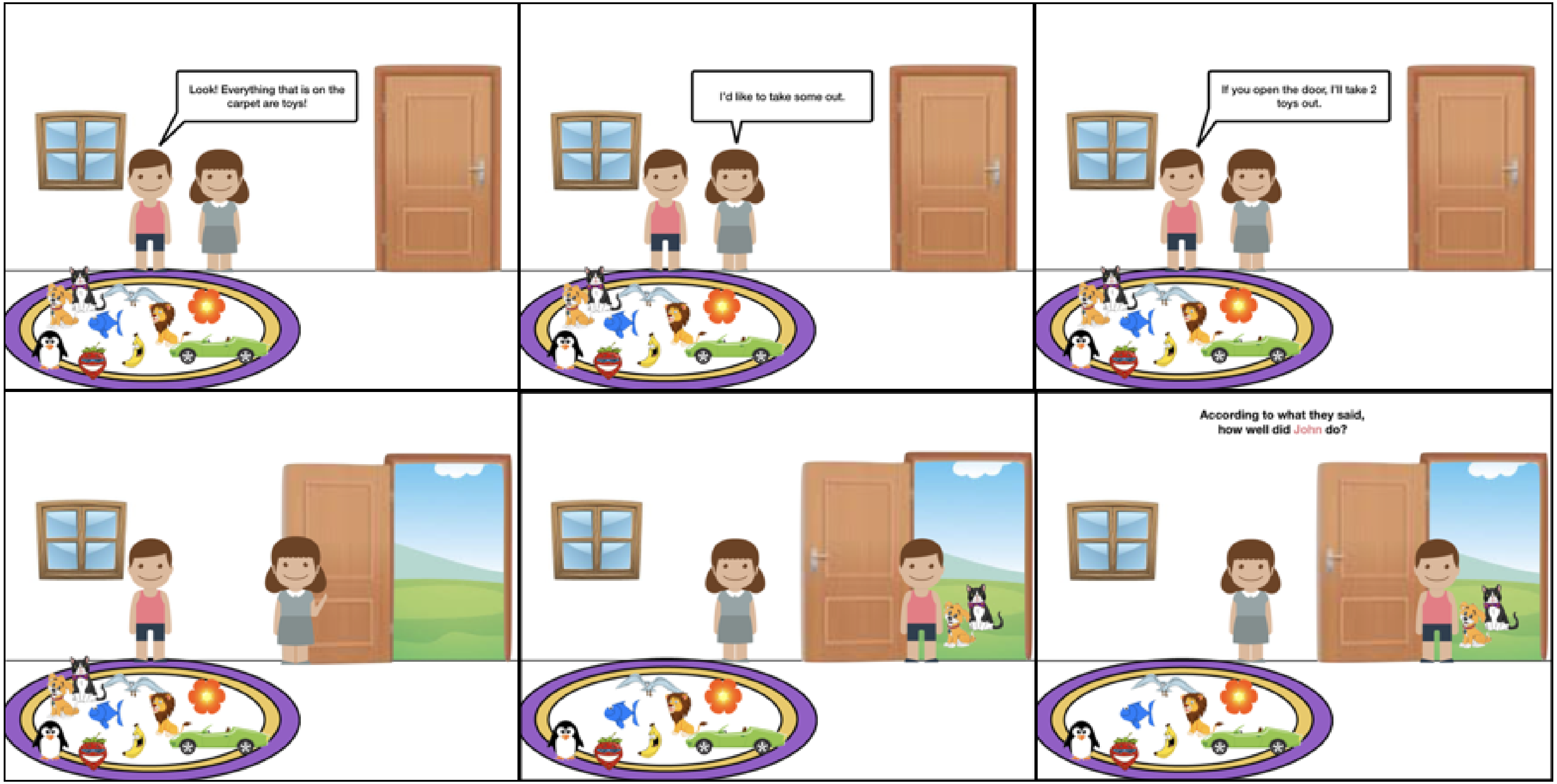

Look! Everything that is on the carpet are toys!
I'd like to take some out.
If you open the door, I'll take 2 toys out.
According to what they said, how well did John do?

Figure 3.2: Example of a test item. First one of the characters (randomized through the experiment) attracted the participant's attention to the objects and then the other character specified his/her goal to manipulate the objects (in that case, to take some objects out). The first character then uttered the target sentence, in that case containing the logical operator 'and' in upward context. After that, the characters performed the actions (in that trial, first the non-actor, and then the actor). The non-actor always performed the action as stated before, and for that-test trial, the actor manipulated the exact number of objects required (exact condition). Finally, the participant was asked to grade the actor character's action (the boy) on a 5-point scale. The role of actor and non-actor was balanced between the boy and the girl characters.

3.2.2 Results

After excluding the two participants who made more than the 20% erroneous answers to the filler items, we analized our results.

Participants rarely used the 'I don't know' button (4.7%), indicating that they were attentive to the stories and that the middle option of the scale could be interpreted as a genuine indication of middle acceptance. We also analyzed fillers and controls. Performance was generally high (mean of erred fillers per participant: 6%; mean of erred controls per participant: 10%).

After this sanity check, we turned to the analysis of test trials. There are two ways to look at the data and classify participants' responses. One way is *theory-neutral*, and it consists in classifying responses according to whether the actor gave less, exactly, or more objects with respect to the content of the embedded clause.

The second way is *semantically inspired*, and considers the meaning of such conditions in the context of the relevant operators, according to whether they suggest an at least or an enriched meaning, or else an interpretation incompatible with any of them. The two classification methods are not equivalent for all connectives: for a disjunction, a high evaluation of a 'Less' story indicates the acceptance of an at least meaning; however, for a number or a conjunction it would correspond to accepting a numerical error. By contrast, a high evaluation for 'Exact' would amount to an endorsement of the at least meaning for disjunctions, but of the enriched meaning for conjunctive and numerical expressions. Finally, a high evaluation for 'More' would indicate an at least reading of all the three kinds of operators. Below, we present analyses for both classifications in turn, starting from the theory-neutral classification.

Due to their ordinal nature, we explored the data with cumulative link models, with the R package Ordinal (Christensen, 2019). The dependent variable was the Median response value per participant per condition. Quantity, Logical operator, Context and their double interactions were predictors (the analysis of the triple interaction was not possible as the model would not converge; see below). The Analysis of Deviance revealed a main effect of Quantity (Wald χ(2, 29)= 92.7185, p<..001) and Logical operator (Wald's χ (2, 29)= 86.4, p<..001) but no effect of Context.

All double interactions were significant predictors (Context x Logical Operators: χ (2, 28)= 6.8, p=.03; Logical Operator x Quantity: χ(4, 26)= 234, p<.001; Con-

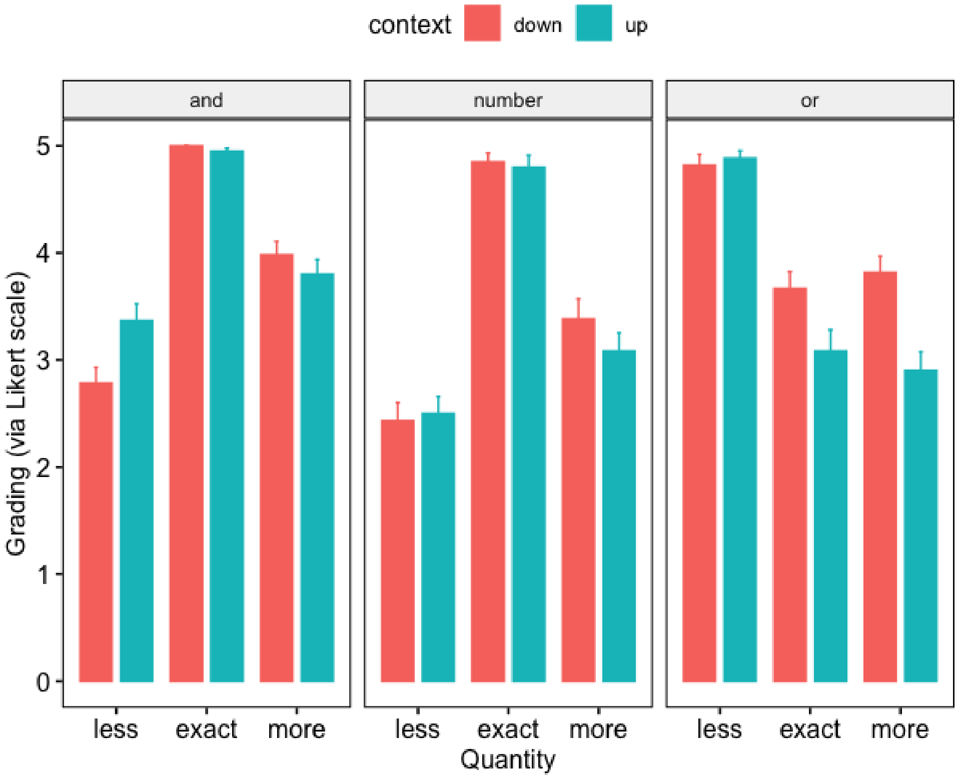


Figure 3.3: General comparisons of UE vs DE results for all quantities and logical operators. Error bars indicate standard error.

text x Quantity: χ(2 26)=9.9, p<.007). These results were confirmed even after adjusting the model for scale effects and proportional odds violations of Logical Operator.

Table 3.3 summarizes the mean, median and standard deviation for each experimental condition combination (Logical operator x Quantity x Context). These results are easier to appreciate by looking at figs. 3.4 and 3.3, which also summarize these results.

Logical operator	**Quantity (theory-neutral)**	**Quantity (semantically motivated)**	**Context**	**Median**	**Mean**	**SD**
and	**exact**	**Enriched**	**DE**	5	5	0

and	**exact**	**Enriched**	**UE**	5	4.95	0.153
and	**less**	**Error**	**DE**	3	2.78	0.806
and	**less**	**Error**	**UE**	4	3.37	0.860
and	**more**	**at least**	**DE**	4	3.98	0.676
and	**more**	**at least**	**UE**	4	3.8	0.750
number	**exact**	**Enriched**	**DE**	5	4.85	0.458
number	**exact**	**Enriched**	**UE**	5	4.8	0.610
number	**less**	**Error**	**DE**	2.25	2.43	0.917
number	**less**	**Error**	**UE**	2	2.5	0.871
number	**more**	**at least**	**DE**	4	3.38	1.02
number	**more**	**at least**	**UE**	3	3.08	0.938
or	**exact**	**at least**	**DE**	3.5	3.67	0.864
or	**exact**	**at least**	**UE**	3	3.08	1.09
or	**less**	**Enriched**	**DE**	5	4.82	0.565
or	**less**	**Enriched**	**UE**	5	4.88	0.387
or	**more**	**at least**	**DE**	4	3.82	0.835
or	**more**	**at least**	**UE**	3	2.9	0.960

Table 3.3: Median, mean and standard derivation per each combination of Logical Operator x Quantity (in the two ways to analize it, theory-neutral and semantically-motivated) X Context.

The analysis shows that participants were sensitive to the specific logical operator in the sentences (main effects of Logical Operator, Quantity, and its interaction), that they interpreted it as a function of its position in the sentence (main effect of Context, and Logical operator x Context interaction), and that the entailment context affected the propensity to accept which quantity was considered to better reflect the interpretation of the conditional sentences (Context x Quantity interaction).

The role of the logical operators in the above results results clearly asks for inspecting participants' responses as a function of their semantic interpretation. Furthermore, inspecting the data, it appears that in some conditions participants selected the highest level of the scale, and that there was no variability, and hence, no information. This fact is interesting in itself, but it also makes data analysis difficult, especially because the data classification, which does not consider the meaning of the connectives involved, masks the nature of the effects.

Thus, because the Quantity levels mean different things for the different logical

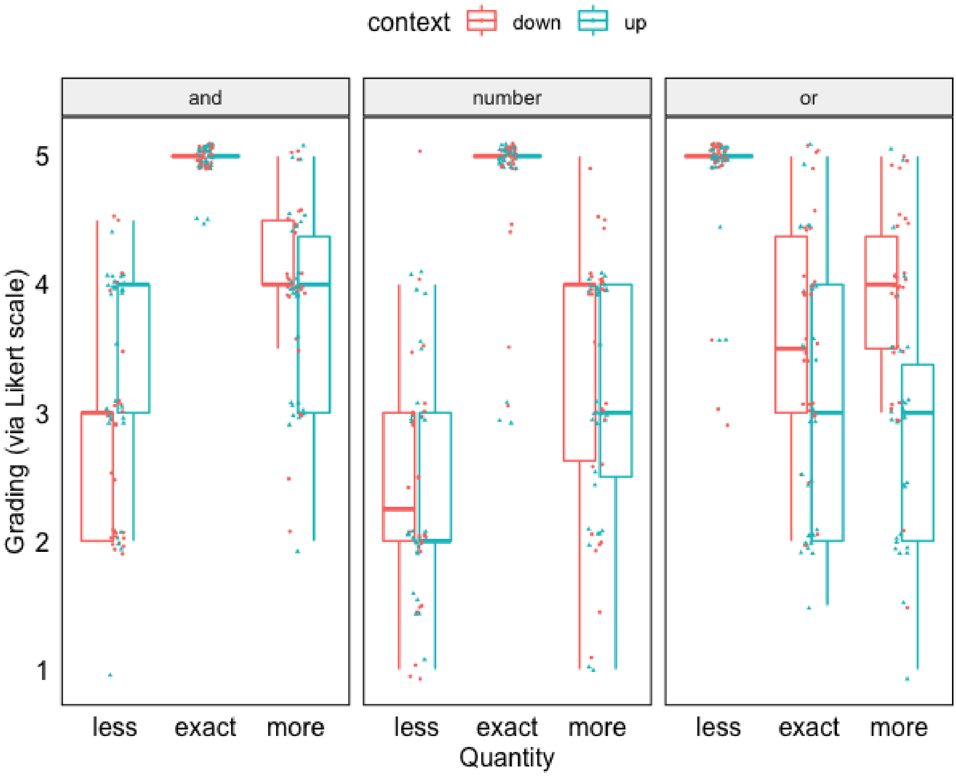


Figure 3.4: Boxplots of the general comparisons of UE vs DE results for all quantities and logical operators. The thick line indicates the median.

operators, we recoded Quantity according to the operator considered. Specifically, for Number and And, ‘Enriched’ corresponded to the ‘Enriched’ interpretation and More to the at least one, whereas Less was coded as an error (but see below). For Or, Less was recoded as an ‘Enriched’ interpretation, while More and ‘Enriched’ were mapped onto an Extended interpretation; unlike And and Number, none of the Quantity conditions could be considered an error, as the only real error for a disjunction would be to ask, say, for A or B and be given only C, and not a subset of, a superset of, or the exact members of the set composed by A and B. Thus, this cell remained empty (see fig. 3.5).

After recoding, it appeared clear that it was the enriched interpretation, when proposed, the one without any variability. It was considered by far the most acceptable, almost without exceptions, regardless of its context (see table 3.3 for the medians of each condition, and also figs. 3.5 and 3.6 for a visual inspection of the data). This aspect of the data clarified, we focused on the remaining contrasts:

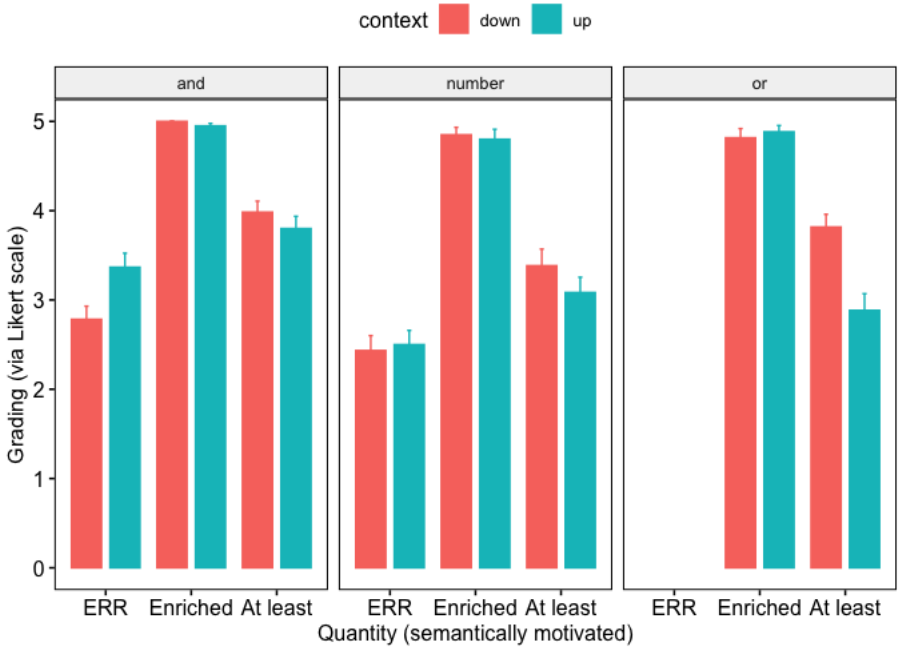


Figure 3.5: General comparisons of UE vs DE results for all semantically-motivated quantities and logical operators. Error bars indicate standard error.

Error and at least for And and Number, and at least for Or.

We begin by analyzing how the acceptance of an at least interpretation varies for each connective as a function of the U/D context. For each operator, we compared participants' median responses to U/D contexts with nonparametric tests, Bonferroni-corrected for multiple comparisons (Exact Wilcoxon-Mann-Whitney Test, R package Coin). For OR, participants considered the at least interpretation as more acceptable in D than in U contexts (Z = 3.6, p<.001, paired, one-tailed), but not so for And or Number, for which the responses showed little sensitivity to the embedding context (respectively, Z = 1.7268, p>0.25 and Z = 1.2, p>0.6).

Our data allowed us to estimate relative differences in preferences between logical interpretations: indeed, accepting an exact interpretation is not necessarily incompatible with accepting an at least interpretation. That said, the distance between degrees of acceptability could be used to determine if the two are indeed considered incompatible, as it would be the case if the former were considered perfectly fit and the latter totally unfit. This does not occur in our data, which suggest that 'Enriched' is always preferred, but also 'at least' may be accepted in the conditions described above.

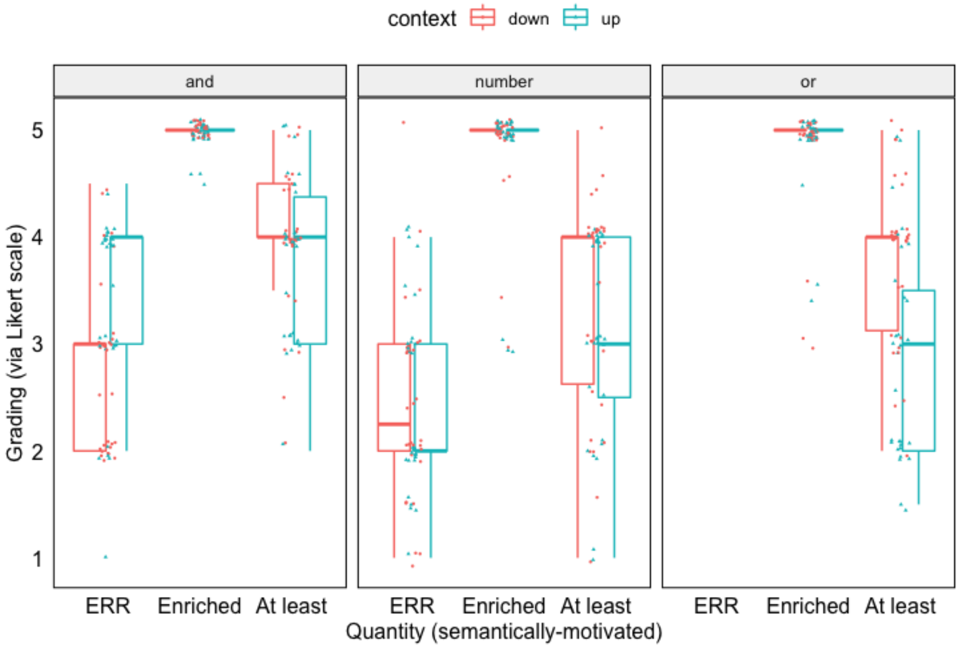


Figure 3.6: Boxplots of the general comparisons of UE vs DE results for all semantically-motivated and logical operators. The thick line indicates the median. Notice that the enriched quantity showed a ceiling effect, with almost no variability in any condition.

By using a truth value task, Panizza, Chierchia, and Clifton Jr (2009) found that participant accept an interpretation of numerals x as ‘at least, x’ (x or more) more frequently in upward than in downward contexts. Our data on the Number condition seem incompatible with this result, but we tested relative judgments and specifically probed the differences between ‘Enriched’ and ‘at least’ that in Panizza, Chierchia, and Clifton Jr (2009).’s design were grouped in an ‘at least’ response. To better compare our result with theirs, we transformed responses to Number to binary responses alike those of a truth value task by recoding the lower end of the scale (1 and 2) to ‘Rejected’ and the highest end (4 and 5) to ‘Accept’, removing responses at the middle of the scale (3) from analysis (2.9% of all responses), and we collapsed ‘Enriched’ and ‘at least’ conditions to an ‘inclusive’ meaning. With this reconstruction, even in our data the ‘inclusive’ meaning was more accepted in D than in U contexts(respectively, (M=84.35% and M=76.27%; one-tailed Exact Wilcoxon-Mann-Whitney Test; Z=1.69, p=.049). Thus, even though the effect was weaker than in Panizza, Chierchia, and Clifton Jr (2009), we conclude that there is no fundamental difference between the two results, but that they evaluate in a different way the points of balance created by the forces of entailment on the one hand and exactness of the meaning of number on the other hand exert against each others.

A second aspect of the data worth highlighting can be seen in participants' responses we classified as Errors in And and Number. In that cases clearly U/D contexts modulated judgments of acceptability. For Numbers, Context had no effect (Exact Wilcoxon-Mann-Whitney Test, p> 0.74). However, for And it had a strong influence (p<.01): in that case participants ranked as more acceptable 'errors' in UE than in DE contexts; furthermore, for And Context interacted with Quantity, as participants inverted their relative judgments in Err and in at least (Exact Wilcoxon-Mann-Whitney Test on the medians of the differences DE-UE Z=-2.6, p<.02), but not for Number (p>.2). This suggests a considerable tolerance for errors specifically with 'and', more pronounced in UE contexts. We comment on these results below.

3.2.3 discussion

Overwhelmingly, participants judged the exact interpretation as the favored one for all three operators. These clear preferences, however, should not obscure other important aspects of the data.

First, to a considerable degree participants tended to accept even the other available options not reducible to an exact interpretation. In particular, our analyses confirm the prediction of the theory that the entailment context (UE/DE) modulates the meaning participants attribute to 'or', toggling it from exclusive in UE to a more pronounced inclusive reading in DE contexts.

Unlike Panizza, Chierchia, and Clifton Jr (2009), we did not find any clear indication that also numerals change their interpretation as a function of the entailment context. However, when the data were treated as in showed that differences in methodology and analysis could explain the discrepancies perhaps suggesting that the effect of entailment context for numbers is weaker than for a logical connective such as 'or' (see Panizza, Huang, Chierchia, and Snedeker, 2009 for more results pointing towards this direction).

Overall, the responses for 'and' and numerals followed a similar pattern. An striking fact assimilating them was also that the 'less' condition, which in principle should be considered an error, was fairly well accepted (for 'and', M=3.075, SD=.87; and for 'numeral', M=2.46, SD=.89). That is surprising, as both constitute a violation of the logical meaning: it would amount to accepting '2' (or an apple and a pear) as a correct response when asked for 3 (or an apple, a pear and a plum), but obviously participants know that 2 is not 3. This suggests that their responses were not uniquely driven by the logical meaning of the expressions. We interpret the results as indicating that when participants interpret a text, they run a double task; first, they locate and interpret the meaning of the logical operators involved in it (and hence effects such as the modulation of the interpretation of 'Or' as a function of the entailment context). Then, however, they evaluate a metric of partial satisfaction of the given conditions, which drives their acceptability. Thus, even though the character did not get 'entirely' what s/he was asked for – say 3 objects – getting two is still better than getting none of the items. If that were true, these judgements should be sensitive to a distance effect: the further away from the number of required items, the worse participants should evaluate the action.

Indeed a similar phenomenon was found by Chemla and Spector (2011). In their task, which involved a continuous scale, subjects tended to rate the relevant sentences not only by evaluating their truth conditions but by taking also into account some 'metric that reflects the distance between a particular situation and some prototypical situation determined by the relevant sequence' (Chemla & Spector, 2011). What our data add to this potential explanation is that, *even in this case*, in which pragmatic considerations drive participants' acceptability judgments, the entailment context plays a fundamental role: our data show that at a parity of violation, participants judge conjunctive sentences (but not numerals) more acceptable when the violation is in an UE context (or in the consequent of a conditional) than in a DE context.

To better explore this interpretation, and in particular, the potential presence of a distance effect, we conducted experiment 2. Because this phenomenon only concerns 'and' and 'numerals', we restricted our study to these two operators.

3.3 Experiment 2

The aim of this experiment was to test the presence of a 'partial satisfaction' strategy that would lead to a graded acceptance of a violation of the meaning of a conjunctive or a numeric sentence as a function of the distance from its actual meaning ('distance effect').

For 'partial satisfaction' we refer to a phenomenon according to which even if the request is not totally satisfied, the action is evaluated as acceptable due to the fact that the request was at least partially satisfied (i.e., one a character is asked to take out three toys, but he/she takes out two, this action is still acceptable since it is still better than taking out zero toys).

For ‘distant effect’ we refer to the fact that an action gets evaluated more negatively the more distant it is from the expected exact meaning (i.e., if one character is asked to get out three toys, the action will be rated lower by the participant in case he/she takes out one rather than in the case he/she takes out two)

On the basis of the ‘partial satisfaction’ and the ‘distant effect’ phenomena, we hypothesised that participants would be partially satisfied with incomplete and overcomplete answers and that the more distant from the exact quantity, the lower the action will be rated .

3.3.1 Methods

Participants

31 Catalan-speaking adults were tested, and one of them was excluded from the analysis due to failure in the filler part of the experiment, where threshold inclusion was under <20

They were recruited from the Pompeu Fabra university database from Barcelona and its surroundings, and given €10 for participation.

Procedure

The procedure was the same as Experiment 1 (see section 3.2): participants saw a story in which one of the characters announced (with a Quantityal sentence) what they would do, and, once the story was over, were asked to grade using a scale from 1 to 5 their actions As in Experiment 1, we introduced an ‘I don’t know’ option, to make sure participants’ responses would not be due to a memory error. In this experiment,, because the number of trials increased with respect to Experiment 1, we introduced a pause at the middle of the study, to ensure participants stayed focused and to give them a break from the experiment. The length of the experiment was about 45 min.

Material

Differently from Experiment 1, only the logical operator ‘and’ and ‘numerals’ were tested (in other words, ‘or’ was not included). The design was very similar to the one used in Experiment 1. There were 92 trials, 32 per logical operator and 28 fillers, half of which had a Quantityal sentence. The main change compared to Experiment 1 was that this time there were 8 different Quantities per logical operator:

- ‘Exact’: the actor manipulated the exact number of objects mentioned in the target sentence, as in Experiment 1.
- ‘n+1’: the actor manipulated one object more than mentioned in the target sentence; this Quantity was the ‘more’ Quantity from Experiment 1.
- ‘n+2: the actor manipulated 2 objects more than mentioned in the target sentence; this Quantity was not present in Experiment 1.
- ‘n+3’: , the actor manipulated 3 objects more than mentioned in the target sentence; this Quantity was not present in Experiment 1.
- ‘n-1’: the actor manipulated one object less than mentioned in the target sentence; this Quantity corresponds to the ‘less’ Quantity of Experiment 1.
- ‘n-2’: the actor manipulated 2 objects less than mentioned in the target sentence; this Quantity was not present in Experiment 1.’n-3’: the actor manipulated 3 objects less than mentioned in the target sentence; this Quantity was not present in Experiment 1,.
- Control sentences: as in Experiment 1, in these sentences the Quantityal was not satisfied, either because the character acting on the target part of the sentence did not provide any item or because the other character did not perform the task (e.g. when told ‘If you open the door, I’ll take 2 toys out’ the character supposed to take the toys would not take any).

As we also manipulated the entailment context (DE versus UE as in Experiment 1), that gave a total of 64 test trials (32 per logical operator, of which in the half of the cases, 16, the logical operator appeared in DE, and the other half in UE. These 16 trials per logical operator and context were obtained by repeating each of the 8 Quantities twice).

3.3.2 Results

We first performed a 3-way repeated-measures ANOVA with the data (Quantity x Logical operator x Context). It showed a main effect of Quantity (F(6,29)=133.56, p $<$.001, η_p^2=.82) and Logical operator (F(1,29)=63.18, p $<$.001, η_p^2=.69), a 2-way interaction between Quantity and Logical operator (F(6,29)=9.93, p $<$.001, η_p^2=.26), between Logical operator and Context (F(1, 29)=10.32, p=.02, η_p^2=.26), and between Quantity and Context (F(1,29)=2.55, p=.02, η_p^2=.08),as well as a 3-way interaction between Quantity, Logical operator and Context (F(6, 29)=3.65, p=.002, η_p^2=.11). We decided to further analyze the data by splitting it by Logical operator and performing 2-way repeated-measures ANOVAs.

For 'numeral', we found a main effect of Quantity (F(6,29)=102.64, p $<$.001, η_p^2=.78), and a tendency for Context (F(1,29)=3.91, p=.057). For 'and' there was a main effect of Quantity (F(6,29)= 121.04, p $<$.001, η_p^2=.81) and of Context (F(1, 29)=5.98, p=.02, η_p^2=.17). and an interaction between Quantity and Context (F(6, 29)= 4.22, p $<$.0001, η_p^2=.13).

Further analyzes using paired t test with Bonferroni correction revealed that Context was statistically significant only in the Quantity 'n-1' (t(29)=4.03, p=0.03; upward context, M=3.03, SD= .76, downward context, M=2.48, SD=.68). Results, split by Logical operator and Context, can be seen in figure **??**.

With these results in mind, we performed a correlation between the absolute magnitude of condition (±1-±3), splitting by Logical operator, as results showed each one worked differently. We used the Rmcorr R package for within- participants correlation.There was a strong negative correlation between sentence evaluation and the magnitude of the violation in 'and' (repeated-measures correlation: R$_RM$(59)= -.70, R2= 0.48, 95% CI= [-.81, -.54], p $<$.001) and in 'numeral' (repeated-measures correlation R$_RM$(59)=-.69, R2= .47, 95% CI= [-.8, -.52], p $<$.001). See figure **??** for more details

3.3.3 Discussion

The results showed that in this setup the evaluation of the response depends on the magnitude of the deviation from the request: the closer the outcome is to the expected outcome, the higher the acceptance is (n-1 and n+1 were more accepted than n-2 and n+2, which were, in turn, more accepted than n-3 and n+3, see Figure 4). That also shows that participants were partially satisfied with answers that failed to satisfy the logical meaning of the sentence under evaluation.

Context (UE vs DE) in this experiment played a smaller role than in Experiment 1. However, it is worth commenting that, as we found in the previous experiment, the Quantity 'less' n-1 for 'and' is better accepted in UE contexts. This result will be addressed in the general discussion.

3.4 General Discussion

The nature of Scalar Implicatures has been long debated, with some arguing they need a certain level of logical reasoning, while some other denying it. In this paper, we try to make a case for the former view by systematically addressing the interpretation of conjunction, disjunction, and numerals embedded in Upward and Downward contexts of a conditional sentence.

Our first experiment confirms that the entailment context (UE/DE) modulates the understanding of 'or' in the way the theory predicts: the inclusive interpretation of 'or' is more accepted in downward entailing contexts. As for numerals, our predictions were only partially confirmed. Contrarily to previous studies looking at entailment effect on the interpretation of numerals, finding a preference for 'exact' interpretation in UE contexts than in DE contexts, studies (Panizza, Chierchia, & Clifton Jr, 2009), we did not find a clear effect. However, multiple articles have shown numerals to work differently than other words with logical valence, such as 'or' (or 'some'). For example, Marty, Chemla, and Spector (2013) found that even though the 'at least' meaning of numerals could be accessed; it came with an extra memory cost associated which is similar to the ones associated with the enriched meaning of 'some'. In fact, although it was clear that the preferred meaning for the 'numerals' was the exact-enriched one, our results show that this 'at least' meaning was accessible by our participants and sensitive to the entailment context .

Another unexpected, although extremely interesting, result from Experiment 1 was that the interpretation of the connective 'and' was also affected by the entailment context: an incomplete, 'less' meaning was judged more positively in upward than in downward contexts. In other words, when three objects were mentioned, participants were happier to see that two were manipulated when the numeral was in the consequent (UE) than in the antecedent (DE) of a conditional — that is, 'if you open the door, I'll take the cat, the dog, and the fish out' was better than 'if you take the cat, the dog, and the fish on the table, I'll close the door' when only the cat and the dog were moved. These results were extremely consistent: we found the same pattern in experiments 1 and 2. We hypothesize this can be due to an asymmetrical application of pragmatic concepts such as partial satisfaction.

We studied the concept of ‘partial satisfaction’ in more detail in Experiment 2. The rationale behind the study was clear: we wanted to be able to target the heuristic forces that were interfering with the experiment. Specifically, we showed that, in this setup, the evaluation of the response was a function of the magnitude of deviation from the request: the closer the outcome to the expected response, the better it was accepted.

Therefore, even if participants were aware of the importance of the logicality in the evaluation of such sentences (Experiment 1), they were also optimizing these evaluations according to the available context (Experiment 2).

It was interesting that, in the ‘less’ Quantity in ‘and’ in Experiment 2, participants showed the same pattern of evaluation as in Experiment 1. That led us to tentatively propose that the effect of partial satisfaction is asymmetrically applied to Upward Entailing.

This idea is supported by the fact that, under these Quantitys, UE and DE contexts carry different entailments: when the conjunction is in the consequent of the conditional (in UE) the sentence entails alternative sentences with fewer conjuncts:

$(\mathrm{R}{\rightarrow}\mathrm{A} \wedge \mathrm{B}) \vDash (\mathrm{R} \rightarrow \mathrm{A})$

That is, in a context in which ‘if you open the door, I’ll take the cat and the dog out’ is true, the sentence ‘if you open the door, I’ll take the cat out’ is also true. In this context, therefore, the presence of ‘and’ will carry no extra implicature. However, the same is not true for DE, that is, when the conjunction is in the antecedent of the conditional:

$(\mathrm{A}{\wedge}\mathrm{B} \rightarrow \mathrm{R}) \nvDash (\mathrm{A}{\rightarrow} \mathrm{R})$

That would be equivalent to say that a situation in which ‘If you take the cat and the dog out, I’ll close the door’ does not tell us anything to what will happen if you only take the cat out. Therefore, that context is likely to trigger the following scalar implicature: NOT (ONLY B $\rightarrow$ R) and NOT (ONLY A $\rightarrow$ R). In other words: for me to close the door, you will need to take both animals out, not just one.

This extra implicature, only present in DE, would prevent participants to apply the same optimization process of being partially satisfied that is available to them in UE and other words such as numerals.

If that explanation were true, we would still need to say some more words about the relationship we scratched between numerals and conjunction. Right now, it seems clear that our participants did not treat them in the same way: the restrictive meaning of numeral (the exact meaning) was always favored respective other interpretations for numerals and respective the conjunction. It seems, therefore, that numerals are interpreted as much more ‘exact’, and that could also explain the difference between ‘and’ and ‘numerals’ in the effects due to partial satisfaction: the exact meaning of the numerals overrides any potential effect of entailment context. In that case, if we put a numeral under our UE context: ‘If you open the door I’ll take three animals out’ would not make ‘if you open the door I’ll take two animals out’ true; for ‘two’ would mean ‘only two’ and that is incompatible with ‘three’. The same would apply to the DE context.

What, then, is the difference between ‘and’ and ‘or’? According to the theory, ‘or’ enriched meaning derives from the negation of stronger alternative terms from the same Horn scale, that is, ‘and’. However, by the same explanation, since ‘and’ is the stronger term of the scale it should not be affected. This is not a problem per se as some studies have shown that it is very easy to construct Scalar Inferences with ad hoc alternatives and that even children are prone to do it (Stiller et al., 2011). That could lead us to suggest that, while the effect shown in ‘or’ is due to the lexical scalar replacement, the effects in ‘and’ would depend on domain, non-grammatical contextual parameters. Indeed, the modulation of ‘or’ due to the entailment pattern was more pervasive than in the case of ‘and’. Interestingly, however, we can see that the effects displayed in ‘and’ and ‘or’ were crossed (‘or’ differed in DE, ‘and’ in UE). And even more interestingly, both results can be explained as a matter of quantity of information depending on the logical properties of the sentence.

In light of the results, we have shown that the effects of logical entailment are more pervasive than previously thought and that they easily interact with other principles of general cognition, such as this optimization of the response based on what is available. Our results show that, in general, language complies with logical principles more systematically.

Chapter 4

Linguistic understanding of entailment

4.1 Introduction

This chapter moves from the study of words with logical valance to entailment, one of the less studied subcomponents of exhaustification. It is assumed to play a role in a multiplicity of cases in the interpretation and the syntactic distribution of linguistic terms; indeed, it is even assumed to have a role in the language acquisition process. Therefore, it is reasonable to expect that entailment relations are accessible and understood early in development.

We will first introduce entailment by offering a summary of the definition and its technical details, and we will continue with a short revision of the empirical results about the understanding of entailment in children. Then, we will present the aims, methods and results of a study on the understanding of entailment relationships in 3- to 6-year-old children and in adults.

4.1.1 What is entailment?

Entailment is a relationship between two sentences such that if the first is true, the second will necessarily be true (Chierchia & McConnell-Ginet, 2000). The notion, borrowed from standard logic, is easier to understand through an example:

(38) a. If Mary takes a ball, she plays with a Dalmatian.
b. If Mary takes a ball, she plays with a dog.

(39) a. If John eats pizza, he gets stains in his t-shirt.
b. If John eats pizza with olives, he gets stains in his t-shirt.

The first example, example (38a), entails example (38b), because a 'Dalmatian' is a dog breed (a subset, in formal terms), and so every time example (38a) is true, so will be (38b). This relationship is not symmetric: example (38b) does not entail example (38a), as not every time 'Mary plays with a dog', she needs to be playing with a 'Dalmatian'.

In a similar way, example (39a) entails example (39b), since 'pizza with olives' is a type (a subset) of pizza, so the fact that John gets stains in his t-shirt when he eats pizza means that when he is eating pizza with olives the stains will also appear. Again, the relationship is asymmetrical: example (39b), does not entail example (39a) (i.e- potentially, the first example could imply that there is something special about 'pizza with olives' that makes John get stains).

Some formal semanticists consider entailment a core property of human languages (Su, Zhou, & Crain, 2012), since all languages contain expressions that restrict their context of appearance according to the entailment context (Ladusaw, 1979; Horn, 1989; Su et al., 2012).

As one can grasp from the previous examples ((38a), (38b), (39a) and (39b)) the entailment relationship established between (38a) and (38b) and between (39a) and (39b) differs. In the first case it extends from a specific to a more general expression: from a set to its superset (that is, from 'Dalmatian' to 'dog'); while in the second case it goes from an expression to something more specific, from a set to its subset (that is, from 'pizza' to 'pizza with olives').

This different relationship depends on the logical context that the target expression (in our case, 'Dalmatian' or 'pizza') is embedded on. It is interesting to stress that entailment is not a property of the sentence as a whole, but it is limited to a specific environment (Gualmini, 2014): examples (38a), (38b), (39a) and (39b) use the same linguistic structure (a conditional: if X, Y). However, the antecedent of the conditional licenses inferences from the set to the subset, while the consequent licenses inferences from the set to the superset.

Contexts like the antecedent of a conditional, in which the expression entails the truth of the sentence in all its more specific expressions (its subsets), as example

(39a) are called downward entailing (DE). By contrast, the consequent of a conditional, in which an expression in a sentence entails the truth of that sentence in all its more general expressions (its supersets), as example (38a), are called Upward entailing contexts (UE) [1].

Entailment contexts, and specifically Downward entailing contexts (DE) are very important for linguistic theory since their existence unifies several apparently disparate linguistic phenomena (Su et al., 2012): words with logical valence interpretation (see chap. 1.2.2 and 3), the distribution of negative and positive polarity items [2] and the non-interrogative use of wh-words in languages such as Mandarin Chinese (Su et al., 2012).

4.1.2 Empirical results about entailment

Even though the importance that the understanding of the entailment relationship carries in linguistic theory, research has focussed on the interpretation of operators within the entailment context, but never on the entailment context *per se*.

For example, different cross-linguistic studies show that children (around 4-5 years-old) are sensitive to the entailment context when determining the meaning of words with logical valence. Thus, Chierchia et al. (2001), Gualmini (2014), Su et al. (2012) investigated the understanding of 'or', while Panizza et al. (2013) focused on the meaning of the numerals, and in all cases they found that children were more prone to give an 'at least' interpretation of the word with logical valence in DE than in UE, in line with the theory.

Other evidence suggests that by 4 years of age children understand the licensing conditions and interpretation of the English negative polarity item 'any', because they use 'any' in DE contexts but refrain to do so in non-DE (O'Leary & Crain, 1994).

An especially interesting result was found by Su et al. (2012). The authors investigated two of the linguistic phenomena that depend on entailment: the interpretation of 'or' and that of wh-words. In their studies, they tested two groups

[1] In this characterization of contexts, we are ignoring non-monotonic expressions (eg. exactly one *X*, more than two *X*, less than 10 *X*)

[2] Negative Polarity Items (NPI) are expressions that are grammatically limited to appear in DE contexts, -eg. 'any', 'ever', 'anymore', 'yet', 'much'. Positive Polarity Items (PPI), on the contrary are ungrammatical in those DE contexts. Examples of PPI are 'several', 'still', 'already', 'some' or 'rather' (Linebarger, 1980; Gualmini, 2014).

of 4-year-old children acquiring Mandarin Chinese. Their results were positive in both contexts: children were aware of the change in interpretation of wh-words in DE (no one) with respect to non-DE (every); likewise, they showed to apply different truth conditions to disjunction depending on the entailment context (Su et al., 2012), accepting the 'at least' meaning of 'or' (see chaps. 1.2.2 and 3 for more details in the duplicity of meanings of 'or') more in DE than in UE. On the basis of these results, the authors concluded that entailment is a core property of language and that children master the different phenomena (the duplicity of meanings of 'or' and the interpretation of wh-words) not item by item, but rather, by relying on the abstract concept of entailment.

Entailment relationships, understood as a tool that children have available in their first years of life, could also play a role in the acquisition of numerals, according to Feiman et al. (2019). The authors investigated the knowledge that preschoolers (3 to 5 year-olds) have of the meaning of their first numeral words when they are in the process of mapping them to quantities (see chapter 2 for a more detailed explanation of such process). They concluded that children understand the entailment relationships that link the numeral words, and use them to advance in their mapping process.

However, as previously stated, these studies do not investigate children's knowledge of entailment relationships directly, but how they can play a role in grasping the meaning of certain operators ('or', numerals, NPIs), usually presented in downward entailing contexts. Therefore, most studies only provide indirect evidence of how children understand entailment relationships.

4.1.3 Aim and hypotheses

Our aim in this set of studies is to directly tackle the question of how and when children are sensitive to abstract entailment patterns in upward entailing (UE) and downward entailing (DE) contexts.

Because an UE context is 'a context in which an expression in a sentence entails the truth of that sentence in all its more general expressions (its supersets)' Horn (1989), Chierchia (2013), we directly tested the hypothesis that children are sensitive to such entailment relation by asking whether they would be willing to draw inferences from the set to the superset (that is, from 'Dalmatian' to 'dog') but not vice versa (so, not from 'dog' to 'Dalmatian'). Inversely, because the definition of DE context is 'a context in which the expression entails the truth of the sentence in all its more specific expressions (its subsets)', we tested whether children would be able to infer from the set to its subsets (that is from 'pizza' to 'pizza with olives') but not vice versa (not from 'pizza with olives' to 'pizza'). Considering their logic, it is thus natural to expect that the pattern of response in DE and UE contexts, should invert :

(40) Expected pattern in UE: from sets to supersets.
  a. From: If John hides inside the tent, he eats pizza with olives.
  b. One infers that: If John hides inside the tent, he eats pizza (superset).
  c. But not: If John hides inside the tent, he eats pizza with olives and mushrooms (subset).

(41) Expected pattern in DE: from sets to subsets.
  a. From: If John eats pizza with olives he gets stains in his t-shirt.
  b. One infers that: If John eats pizza with olives and mushrooms he gets stains in his t-shirt (subset).
  c. But not: I If John eats pizza he gets stains in his t-shirt (superset).

According to the aforementioned literature, we expected our participants to accept what we will now call 'valid inferences' (so inferences from the set to the superset in UE contexts and inferences from the set to the subset in DE contexts) and to refuse invalid ones (that is, inferences from the set to the subset in UE contexts and those from the set to the superset in DE contexts).

We focused on conditional sentences as the linguistic context for testing sensitivity to UE and DE patterns of entailment for two main reasons. In the first place, conditional sentences, which seem to be comprehended already by 3 year olds (Scholnick & Wing, 1995), contain a DE context in their antecedent (the 'if' part of the sentence) and a UE context in their consequent. Therefore, by using the same linguistic material but inverting the order of antecedent and consequence we can obtain a UE or DE context (see the example sentences in (40)and (41)). The second reason has to do with the fact that under certain assumptions 'if' sentences might be translated into non-verbal scenes. Conditionals are often related with causal relationships (Menzies & Beebee, 2020), even in young children (Jorgensen & Falmagne, 1992). Because our long-term aim is to explore the primitive operations underlying the patterns of logical implications involved in understanding linguistic expressions, we could exploit the conditional nature of a causal relation ('if cause then effect') to generate 'nonverbal' UE and DE contexts, and with them, try to investigate how precursors of linguistic implication patterns may be available in preverbal infants watching causal scenes.

Moreover, in order to control for undesired attached meanings to our reference sets, we used compositional concepts (i.e. concepts that were created in a transparent way, such as ‘dog with stains’ instead of ‘Dalmatian’). Furthermore, we hypothesized that such a construction could show the structure of the sets more clearly than using lexicalized concepts.

In what follows, we will first present the results of an adult experiment whose purpose is to tune the stimuli and set a comparison baseline for the study of children’s understanding of UE/DE contexts (experiment 1). Then, experiments 2 and 3 will explore 3-to-6 year-olds’ comprehension of such contexts.

4.2 Experiment 1: Adults understanding of entailment

4.2.1 Methods

Participants

The data from 20 Catalan speakers (M= 22.5, range= 20-25, 14 females) were analyzed in a first version (v.1) of the study. Four additional participants were excluded because they did not meet the minimum criteria of inclusion (see section 4.2.2 for details). Fourteen different Catalan speakers (M= 20.2, range= 18- 23, 7 females) were analyzed in a second version of the study (v.2). One additional participants was excluded for not meeting the minimum criteria. Participants were recruited from the CBC (Center for Brain and Cognition- UPF, barcelona) database and received 5 euros for their participation.

Materials

We prepared four computerized cartoon-like stories using the Apple Keynote software, and transformed into moves of QuickTime File Format by the same program.

In the stories, three friends were picking up different objects. After that, a new friend would come and ask who had taken a particular object. Each of the stories used a different set of objects: pizza, cakes, dogs and trucks. A male voice narrated the stories and made the voices of the characters of the video, in a children-directed speech manner. All the stories begun introducing the main character (which could be a boy or a girl, counterbalanced) and then his/her friends. After that, the narrator specified the children's goals, which varied according to the set of objects that was being used in the story, and showed the two friends selecting one of the objects. Then, the narrator informed the participant that it was the main character's turn, and that his/her actions were function of a conditional sentence such as:

(42) If John enters into the tent, he eats pizza with olives. (Upward context)

(43) If John eats pizza with olives he gets stains in his t-shirt. (Downward context)

As it appears by inspecting the sentences, the target set (in that case, 'pizza with olives') could appear in the antecedent or the consequent of the conditional. After that, the actions of the main character varied depending on the context and condition of the trial. In UE contexts, the main character always overtly verified the antecedent of the conditional (in that case, John would have gone inside the tent). In DE contexts, participants were informed that the main character either acted on the superset or the subset of the mentioned set (so, in that case they were told that John was eating 'pizza' for the Superset Condition, and 'pizza with olives' for the Subset Condition). In a control condition, for both contexts, the antecedent of the conditional was falsified: in our example, for UE Contexts, John would stay outside the tent, while for DE Contexts, John would eat pepperoni pizza. At that point the video would stop and the participant was asked to answer a control question, which required either drawing the inference in UE contexts (in our example, it would ask 'What will John eat?') or else simply recalling the exact object the main character is manipulating ('What is John eating?'), We will call this question the 'Conditional/Memory Question', because it controls whether participants correctly drew the inference or whether they remembered the target object.

Afterwards, the video resumed. The target child would move to a position that would allow him/her to realize the consequent of the conditional sentence, but crucially, s/he was partially hidden, so that participants could never see whether s/he was realizing the action or not (in our example, in an UE context, John would enter into the tent and move the mouth as it was eating, though participants could not see what he was eating, while in DE contexts John was behind a table, so that participants could not see whether he was getting stains or not).

Following that movement, a new friend (a boy or a girl, also counterbalanced) came into the scene, greeted all the others and asked them about their actions. Again, the question depended on the context and condition of each trial. In UE context, the new friend asked who was manipulating either the superset or the subset of the mentioned objects (in our example, the new friend would ask 'who is eating pizza?' in the Superset Condition and 'who is eating pizza with olives and mushrooms?' in the Subset and Control Conditions). In DE Context, s/he asked who performed the actions coded in the consequent of the conditional (in our case, s/he would ask 'who got stains?').

Participants had to respond by pressing the number (1-4) assigned to the child who had performed the action (see figure 1 5.1).

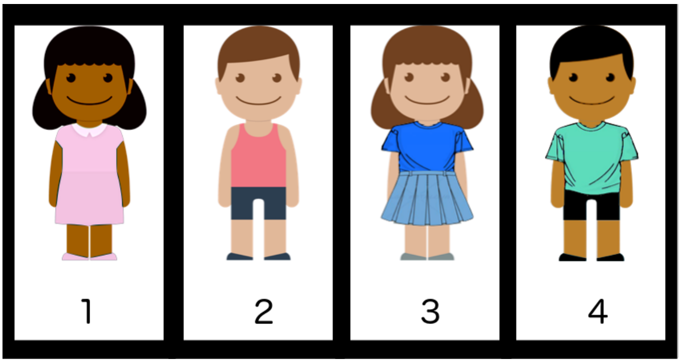
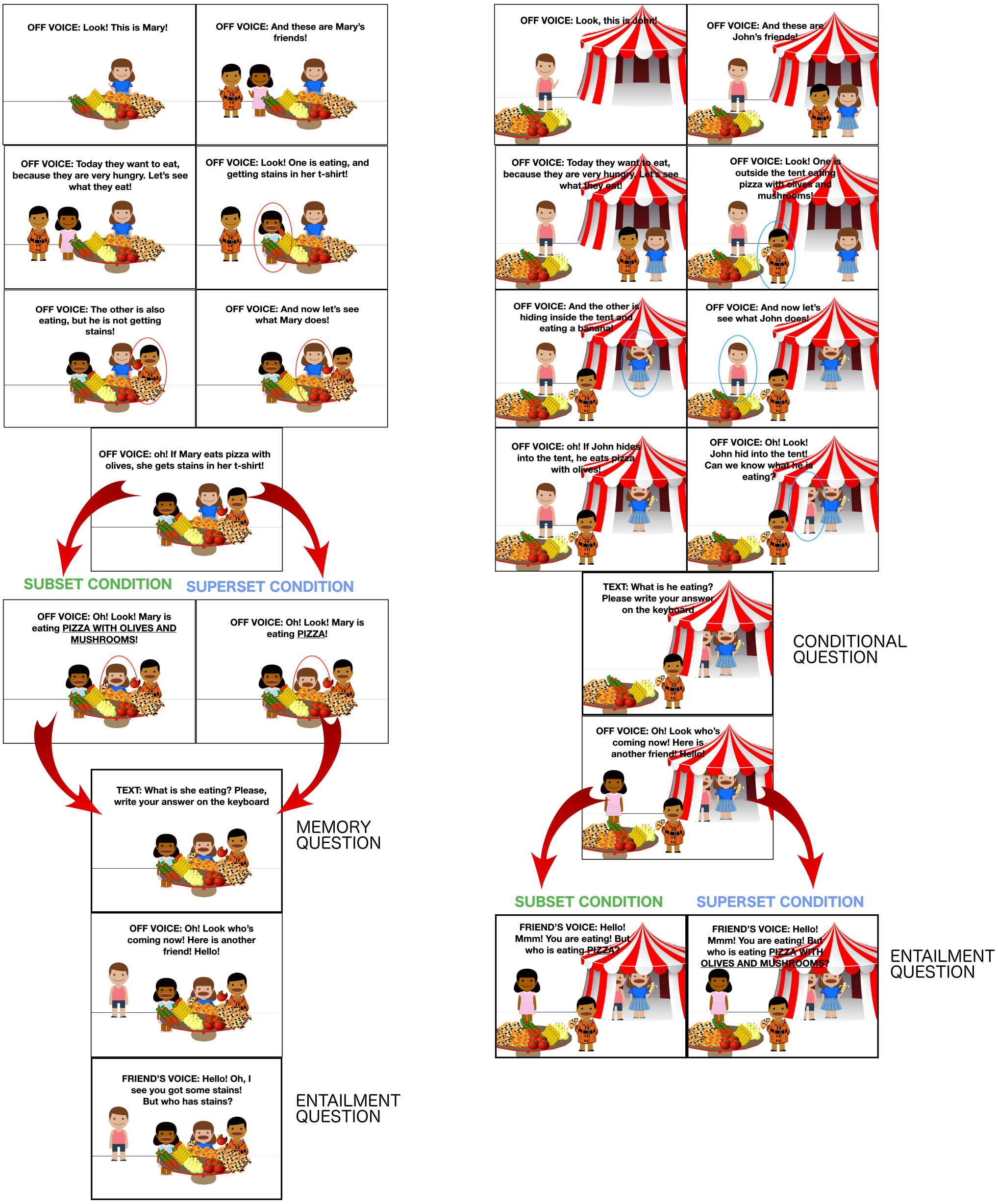


Figure 4.1: Scale of response of test items.

Crucially, one of the non-target characters always fitted the description, and so had to be selected (in our example, for UE context, there was always a child eating pizza with olives and mushrooms, who qualified for both the superset and the subset questions, and in DE context, there was a child that got visible stains while eating). We will call this the 'Entailment Question' because the answer participants had to give depended on the entailment pattern. Figures 4.2a and 4.2b display the structure of the trials.


OFF VOICE: Look! This is Mary!
OFF VOICE: And these are Mary's friends!
OFF VOICE: Today they want to eat, because they are very hungry. Let's see what they eat!
OFF VOICE: Look! One is eating, and getting stains in her t-shirt!
OFF VOICE: The other is also eating, but he is not getting stains!
OFF VOICE: And now let's see what Mary does!
OFF VOICE: oh! If Mary eats pizza with olives, she gets stains in her t-shirt!
SUBSET CONDITION
SUPERSET CONDITION
OFF VOICE: Oh! Look! Mary is eating PIZZA WITH OLIVES AND MUSHROOMS!
OFF VOICE: Oh! Look! Mary is eating PIZZA!
TEXT: What is she eating? Please, write your answer on the keyboard
MEMORY QUESTION
OFF VOICE: Oh! Look who's coming now! Here is another friend! Hello!
FRIEND'S VOICE: Hello! Oh, I see you got some stains! But who has stains?
ENTAILMENT QUESTION
OFF VOICE: Look, this is John!
OFF VOICE: And these are John's friends!
OFF VOICE: Today they want to eat, because they are very hungry. Let's see what they eat!
OFF VOICE: Look! One is outside the tent eating pizza with olives and mushrooms!
OFF VOICE: And the other is hiding inside the tent and eating a banana!
OFF VOICE: And now let's see what John does!
OFF VOICE: oh! If John hides into the tent, he eats pizza with olives!
OFF VOICE: Oh! Look! John hid into the tent! Can we know what he is eating?
TEXT: What is he eating? Please write your answer on the keyboard
CONDITIONAL QUESTION
OFF VOICE: Oh! Look who's coming now! Here is another friend! Hello!
SUBSET CONDITION
SUPERSET CONDITION
FRIEND'S VOICE: Hello! Mmm! You are eating! But who is eating PIZZA?
FRIEND'S VOICE: Hello! Mmm! You are eating! But who is eating PIZZA WITH OLIVES AND MUSHROOMS?
ENTAILMENT QUESTION

(a) Downward entailing trial

(b) Upward entailing trial

Figure 4.2: Example of a Test item for downward context (left) and Upward context (right).

In v.2 but not in v.1, we also created four introductory stories (related to the four test stories) in which two characters, 'mum' and 'dad' were manipulating the objects of the superset (pizzas, cakes, dogs or trucks) and labelling them. Participants saw two of them, counterbalanced. In these stories one of the parents expressed the desire to put all the objects of the superset in a place 'for the kids to play/eat', and then s/he would bring in the target set, specifying the subsets. In our pizza example, mum would bring pizza for the children to eat. She would first bring pepperoni pizza and then pizza with olives, specifying that she was bringing one subset of 'pizza with olives and mushrooms' and another of 'pizza with olives and ham'. Afterwards, the parent would leave and the other parent would arrive, and ask the participant to find the target set. In our example, dad would come and ask where the 'pizza with olives' was. Participants had to respond by pressing the assigned numbers on the keyboard. The experiment would not continue until participants had pressed both the key of the number of the pizza with olives and mushrooms and the one for the pizza with olives and ham (see figure 4.3). Figure 4.4 presents an example of this experimental material.

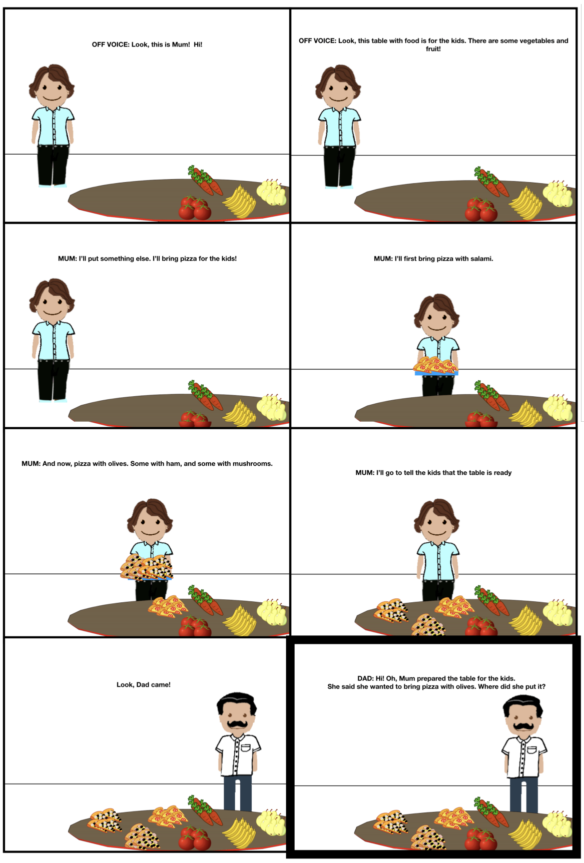

OFF VOICE: Look, this is Mum! Hi!
OFF VOICE: Look, this table with food is for the kids. There are some vegetables and fruit!
MUM: I'll put something else. I'll bring pizza for the kids!
MUM: I'll first bring pizza with salami.
MUM: And now, pizza with olives. Some with ham, and some with mushrooms.
MUM: I'll go to tell the kids that the table is ready
Look, Dad came!
DAD: Hi! Oh, Mum prepared the table for the kids.
She said she wanted to bring pizza with olives. Where did she put it?

Figure 4.3: Example of an introductory story.

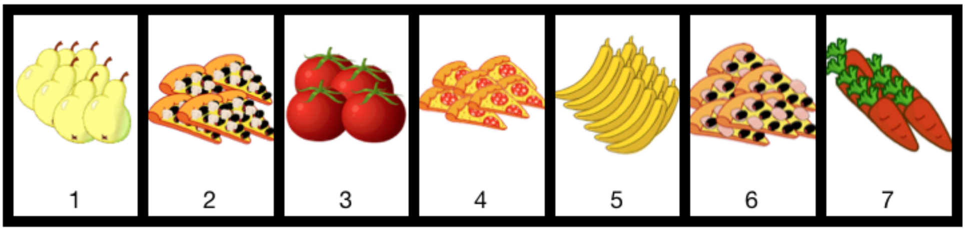


Figure 4.4: Scale of response for the introduction story.

You can find the experimental materials with the English translation of the narrator sentences, in the Supplementary Material folder.

Procedure

Before starting both versions, we informed participants that they would see stories coming from an experiment designed for children, acting as control. This warning was especially useful for them to understand why some questions involved very obvious responses, such as a memory question asking for the action they had immediately seen. The task was administered using PsyScope X B88 (Cohen et al., 1993). Participants were tested in a quiet room in the Center for Brain and Cognition Laboratory (https://www.upf.edu/web/cbc), at Universitat Pompeu Fabra (Barcelona). They wore headphones for the entire experimental session. The study lasted 40 minutes.

The experiment comprised 20 trials, organized in ‘blocks’ of 5 (4 test conditions [UE-Superset, UE-Subset, DE-Superset, DE-Subset] + 1 control [which could be Upward or Downward, depending on the block], order of the trials randomized between blocks), without any pause between trials. Also block order and their structure was randomized, always following the constraint that every block had to contain all four combinations of test conditions (UE-subset, UE-superset, DE-subset and DE-superset) and one control. Thus, the complete design crossed 2 Contexts (Upward vs Downward) x 3 Conditions (Subset vs Superset vs Control), generating 16 test trials (4x4) and 4 controls (2 Upward and 2 Downward per participant).

Summing up, participants had to answer 2 questions per story: a first Memory/Conditional question, asking to resolve the conditional (UE contexts) or to remember what was said (DE contexts); and a final question (the Entailment question) asking them to select all the children for whom a statement was true (always a control child and, depending on the condition, the target child). The selection of the target child was our main dependent variable.

The main difference between v.1 and v.2 of the experiment was that in v.2 the experiment (the Test stories) was preceded by two introduction stories. The introductory stories were added to help participants understand how the test stories developed, and to try to prevent them from being put off by unwanted pragmatical inferences. For example, when listening to ‘pizza with olives and mushrooms’ and ‘pizza with olives’, participants may have concluded that the meaning of ‘pizza with olives’ was ‘pizza with olives *but not mushrooms*’, thus classifying ‘pizza with olives and mushrooms’ and ‘pizza with olives’ as exclusive partitions, instead of seeing ‘pizza with olives and mushrooms’ as a subset of pizzas with olives. Version v.2 controlled for this possible interpretation.

Table 4.1 summarizes our experimental conditions and the expected answer according to the theory:

Context	**Condition**	**Expected selection of target child**
Upward	**Superset**	YES
	Subset	NO
Downward	**Superset**	NO
	Subset	YES

Table 4.1: Summary of the experimental conditions and the expected answer

4.2.2 Results

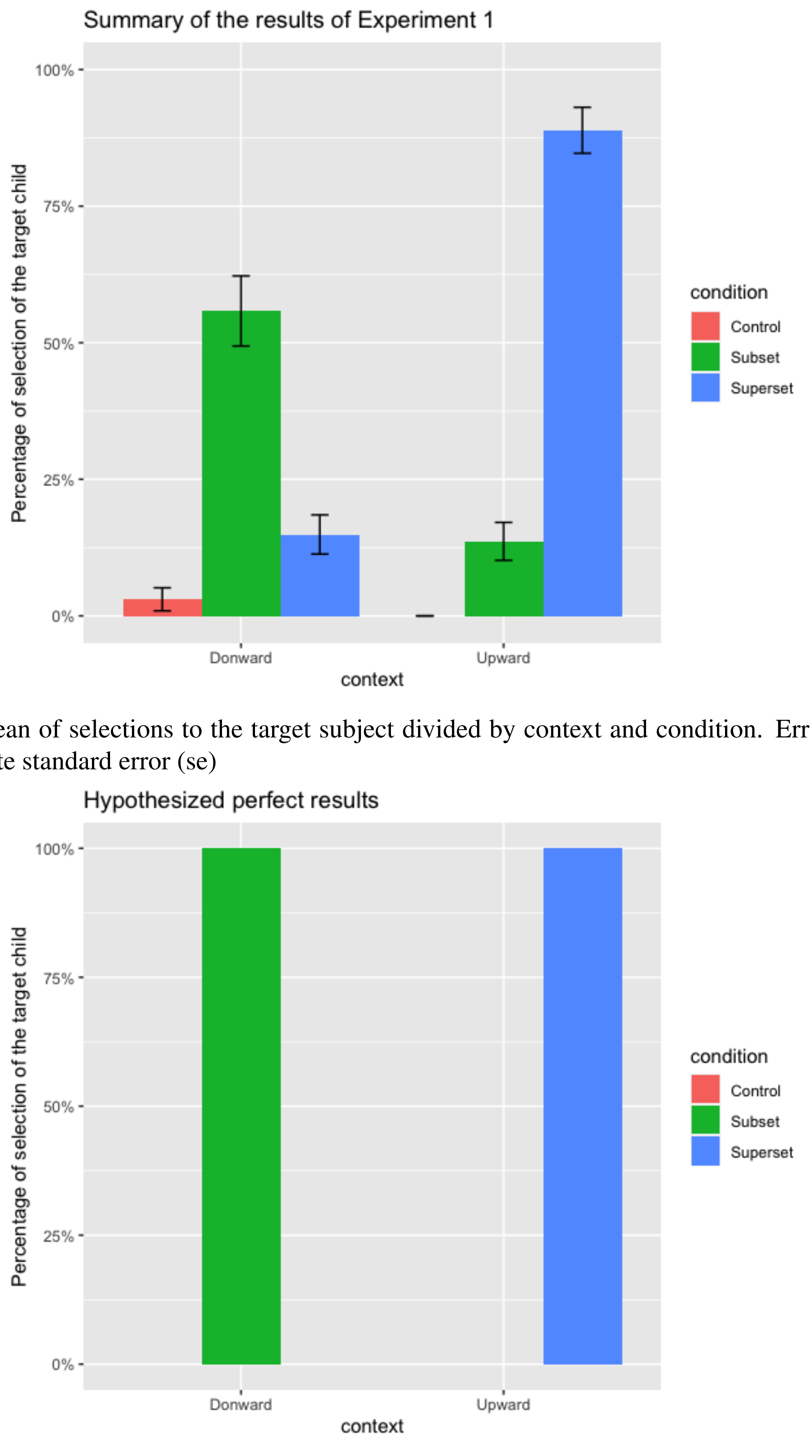
We first filtered the results by using the responses on the Memory/Conditional Question Participants who made more than 25% of errors in these trials (5) were excluded from analysis. Using that criteria, we excluded 4 participants from version 1 but none from version 2.

Of the remaining data, we also excluded the trials in which participants gave incorrect answer to the Memory/Conditional Question (n= 25 in v1; n= 22 in in v2 or the 6.4% of all trials) as well as those trials in which participants failed to select the control child in the Entailment Question (n =8 in v1; n = 6 in v2, or the 1.7% of all trials). By this procedure, one subject from Version 2 was left with no valid trials for experimental condition UE- superset, so we removed her data.

We did not consider an error in the Memory/Conditional Question 8 trials in DE context in which instead of answering about memory (as required in DE contexts), participants draw the inference (as required in UE contexts) nor 10 trials of UE-control condition in which subjects, instead of answering that it was not possible to draw an inference with the information we had given them, answered mentioning any object but the target set. With the remaining data (607 trials from 33 participants, 20 from v.1 and 13 form v.2), we investigated the selections of the target character in the Entailment Question as a function of the context (UE vs DE) and condition (Superset vs Subset).

Preliminary analysis showed no differences between v.1 and v.2, so we collapsed all data and performed a Generalized Linear Mixed Model. The maximal converging model included Context (Upward vs Downward) and Condition (Superset versus Subset) as Fixed effects as well as their interaction. All fixed effects were coded with contrast coding. Participant was included as intercept and by random slope by Context and by Condition; and Story was included as random intercept. We found a significant effect for Context (β = -1.15, se= .32, $\chi^2(1)$ = 12.16, p <.001) as well as for condition (β = -1.25, se= .32, $\chi^2(1)$ = 13.42, p <.001), and an interaction between the two ($\beta = 7.57$, se= .78, $\chi^2(1) = 214.98$, p <.001). Pairwise comparisons using Wilcoxon rank sum test with continuity Bonferroni correction confirmed that invalid inference conditions, that is, UE-subset and DE-superset conditions, were not different from each other (DE- superset: M = .15, sd = .21;UE-subset: M = .15, sd = .22; p=1), but they were different from both UE-superset and DE-subset (all p's<.001). UE-superset and DE-subset, both valid inference conditions, were different from each other (DE-subset: M= .57, sd = .37; UE-superset: M =.89, sd = .24; p=.001). As we predicted, participants selected target character in the control conditions in a very low percentage (DE-control: M = .03; sd= .12; UE-control: M = .0, sd= .0).

Figure 4.5a is a summary of our findings.



(a) Mean of selections to the target subject divided by context and condition. Error bars indicate standard error (se)

(b) Expected results for the experiment

Figure 4.5: Comparison of adults' results with our expected results

While the overall pattern of results was in accord with our predictions, this last result, due to a relatively low level of correct responses in the DE-subset, was unexpected. In order to examine whether participants' initial interpretation of the task could account for this divergence, we further inspected responses in the DE-subset condition. We analyzed potential changes over time of participants' responses by splitting them into four temporal blocks of 5 trials, each of which contained one DE-subset test trial. A trend towards improving correct responses across experiment appeared (see figure 4.6). A Spearman correlation confirmed the visually-salient trend (rank= .29, p=.003)

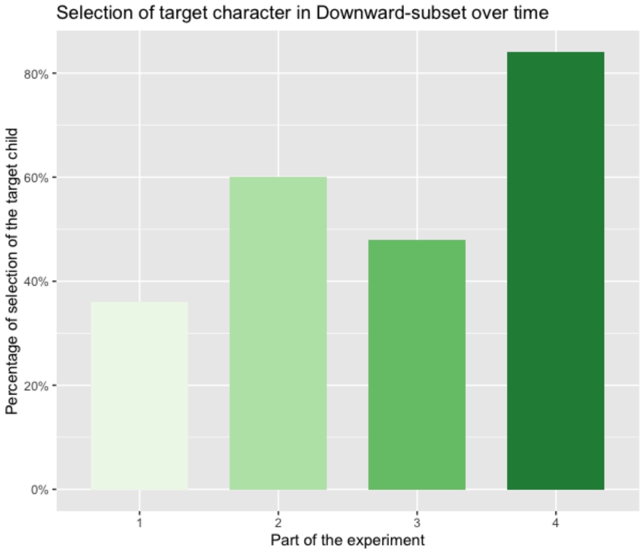


Figure 4.6: Selection of target character in DE-subset experimental condition over the four parts of the experiment. Notice that the number of participants selecting the target character increased over time.

Finally, we looked at individual response patterns, trying to identify the logical type of response patterns. We classified participants (figure 4.7) in the following groups:

1. Validity responders: Participants who selected the target character in valid inference conditions (both UE-superset and DE-subset) at least 50% of the times and at least twice as many times as any other experimental condition. An example of this category is given by participants 16 to 34 figure 4.7.

2. Superset responders: Participants who behaved as above, but in the Superset Condition (that is, they selected the target character in upward-superset at

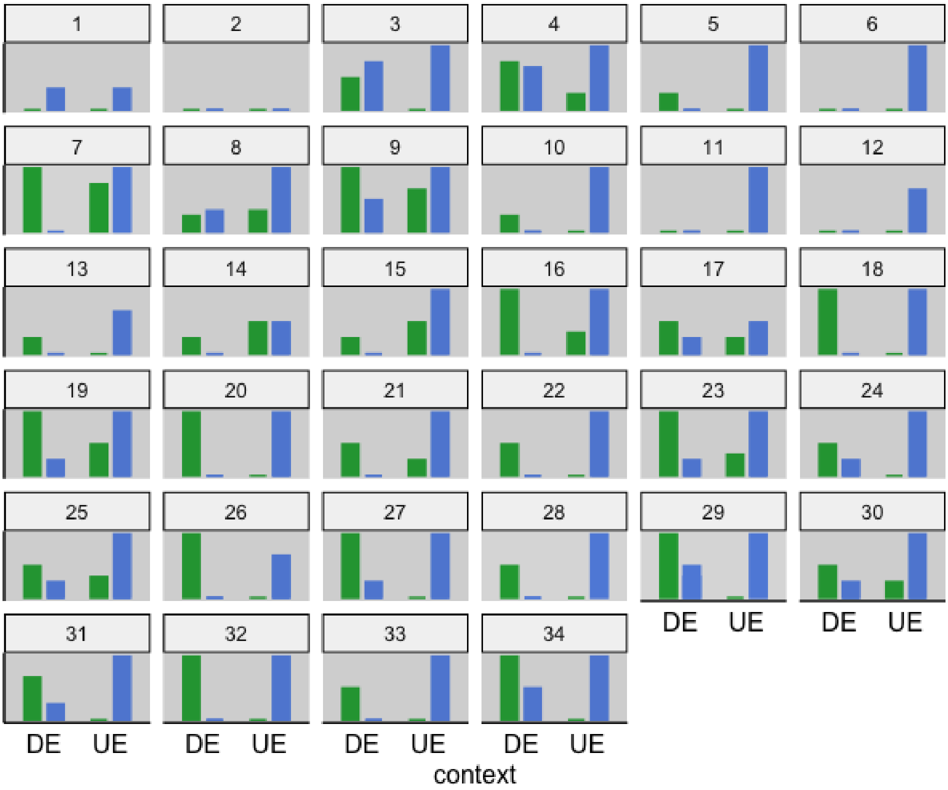


Figure 4.7: Individual adult patterns of responses.

least 50% of the times and at least twice as many times as the downward-superset condition), but drew no difference in Subset conditions (in UE-subset vs downward-subset). Participants 5 to 13 would be examples of this category.

3. Subset responders: Participants who behaved as above, but only in the Subset Condition; Participants 14 and 15 are examples of this pattern of response.

4. Upward responders: Participants who selected the target character in Upward context at least 50% of the times and at least twice as many times as in Downward contexts, thus not showing any modulation of responses as a function of condition. Participants 12 and 18 fall into this category.

5. All-superset responders: Participants who selected the target character in Superset Conditions at least 50% of the times and at least twice as many

Response pattern	Number of subjects
Validity responders	19
Superset responders	8
Subset responders	2
Upward responders	2
All-subset responders	1
None responders	1

Table 4.2: Number of subjects who fell into each response pattern.

times as in the Subset Conditions, thus not showing any modulation as a function of context. Participant 1 is the only example of this category.

6. None responders: Subjects who never selected the target character. Participant 2 was the only one to fall into this category.

Table 4.2 summarizes the number of participants falling into each category. Notice that most participants were classified as validity responders. That is, most participants displayed a pattern of responses coherent to the theoretical expectations. Furthermore over 90% of participants were validity, superset or subset responders, that is, they were sensitive to the UE vs DE contexts. A Fisher's Exact Test for Count Data showed that the distribution of these response patterns was significantly different from random distribution (p=.001). These patterns will be particularly important when we will compare adults with children (see chap. 4.3.2).

4.2.3 Discussion

The aim of this study was to establish a baseline to better understand children's responses when they have to understand situations involving entailment patterns licensed by the same syntactic structures. As expected, participants were sensitive to such entailment patterns, and thus differentiated between conditions in which the invited inference was possible according to the logical context (UE-superset and DE-subset) and those in which it was not possible. Interestingly, inferences in UE-superset contexts were perceived as stronger than those from DE-subset contexts, even though our a priori theoretical analysis would predict an alike effect. However, the responses of DE-subset stories revealed a trend to improve across blocks: participants tended to accept the inference more at the end of the study rather than at the beginning. This effect can be interpreted in different ways. One possible interpretation is that participants do not immediately realize the nature of the task, and that contrast with other similar stories puts in clearer light the proper implication pattern. According to this interpretation, lower acceptance of implication pattens in DE-subset stories is an effect of the pragmatics of the task. Usually, specification of subsets has a contrastive effect (similar to the class inclusion question, see Politzer, 2016), so that mentioning a 'pizza with olives and mushrooms' induces to think of the complementary set of pizzas with olives and something else (or without mushrooms), as an alternative choice to the pizza with mushroom and olives, as opposed to a subset of pizzas with olives. The fact that initial stories aimed at clarifying the subset-set relation holding between relevant objects (v.2) had no effect on performance may indicate that pragmatic alone is not a sufficient explanation. Another (non exclusive) possibility is that DE-subset inferences are less spontaneous and immediate than the UE-superset ones which elicited almost perfect responses, and participants may have to overcome some resistance to draw inferences towards the DE-subset, at least in our task. Whichever the explanation of the lower performance in DE-subset stories, the high level of participants sensitive to valid patterns of implications (the validity responders in our classification) signals that the study can successfully capture their intuitions on the inferences available in each entailment context. With these results in hand, we now turn to children's responses to these experimental materials.

4.3 Experiment 2: Children's understanding of entailment

In this study, we investigated children's understanding of entailment. We did so by using the same experimental materials validated with adults in Experiment 1. We focused on participants as young as 3 year-olds, an age slightly below that on which questions about the understanding of entailment relations had been previously tested (Chierchia et al., 2001; Su et al., 2012; Panizza et al., 2013; Gualmini, 2014). Contrarily to previous research, we focussed on entailment relationships directly, not in how they can play a role in the meaning of certain operators. Our oldest test group was 6, far beyond the age of success in previous studies involving entailment. We tested two different groups of children: Catalan speakers and Italian speakers. Even if previous studies suggest that linguistic experience does not seem to play a major role in children's understanding of phenomena that depend on entailment (Su et al., 2012; Gualmini, 2014), no study exists on the specific patterns of implications we are testing. Thus, a cross-linguistic comparison could allow us to explore whether linguistic differences could affect which kinds of logical relations could be accepted and, potentially, the age at which children would master them. Catalan and Italian are interesting choice because they show different patterns of NPIs (e.g. 'any' exists in Catalan but not in Italian). Unfortunately, the recollection of data for the Italian group is still ongoing, so we will not draw any strong conclusion based on language.

4.3.1 Methods

Participants

119 Catalan children, from 3 to 6 years of age, participated to this study. Additional 23 children were excluded because they did not complete at least 5 trials (out of 10), due to technical or connection problems, or because children refused to participate. Of the remaining 96 children 19 were excluded because they did not meet the minimum criteria of inclusion (see section 4.3.2 for details).

Thus, a sample of 77 children (15 3-year-olds, 18 4 year-olds, 21 5-year-olds and 23 6-year olds; M = 4.6; 39 females) were retained from analysis. They came from Catalonia, most of them from Barcelona and the great Barcelona area (62). Half of them were Catalan speakers (35), the remaining (42) were bilingual Catalan-Spanish (38), Catalan-English (1) or Catalan-French (1) and 3 of them were trilingual (Catalan-Spanish-English). They were recruited among participant families in the Babylab database of the Center for Brain and Cognition or by direct contacts among friends and relatives of the author of the dissertation. They were tested online and sent a small present for participation.

Additionally, at the time of the completion of this dissertation, 28 Italian children were recruited. One was excluded due to auditory problems, and 8 because they did not meet the minimum criteria of inclusion (see section 4.3.2 for details). The remaining 19 children (7 3-year-olds, 5 4 year-olds, 4 5-year-olds and 3 6-year olds; M= 4.26, 8 females), mainly came from Italy, especially from Lombardia (6). 11 were monolinguals in Italian, and 8 were bilingual (French-Italian (4), Ladino-Italian, German-Italian and Slovenian-Italian (2)). They were recruited from personal contacts, and tested online. However, due to logistic issues, they were not sent any participation present. The recruitment and testing of the Italian group is still ongoing, and was limited by the difficulty at contacting participants during the pandemic period and by the slow response of online contacts.

Materials

We used the same experimental material and rationale described in Experiment 1 (see section 4.2.1), with two main differences. The first main difference was that since it was run online, in order to keep children's attention, we created a video of the experimenter commenting the story which was superposed at the right corner of the main video story.

The second main difference was on how the experimental questions were presented to children. Recall that the study asked two questions to participants, one at the middle of each story (the Conditional/Memory Question, in which participants had to type an answer), and one at the end (the Entailment Question, in which participants had to select the characters who had performed the asked action). In this version, children were asked the Conditional/Memory Question orally, and to answer the Entailment Question by touching (if using a smartphone or tablet) or pointing (and then their parents clicking where they pointed, if using a computer) to their selected characters. Both questions were asked by the recorded experimenter.

Procedure

The experiment was run online, using the PCIbex platform (Zehr & Schwarz, 2018), an online platform which allows to easily program and run experiments in Chrome, Edge and Mozilla browsers. Participants' parents or guardians were asked to run the study by their children. At the beginning of the study, parents were informed about the aim and the procedure, and were asked to fill in a questionnaire and record a video as informed consent. In this part they also received precise instructions about their role in the study. We stressed the importance to encourage children to answer but not to correct them.

We restricted the experiments duration to 10 trials plus one introduction trial (which was the half of that of adults). Trials were organized in the same way as in adults: in blocks of 5 (4 test conditions [UE-superset, UE-subset, DE-superset, DE-subset] + 1 control [which could be UE or DE, depending on the block]). The order of the trials was randomized between and within blocks. There were 8 experimental lists.

After participants had finished a trial, the next one automatically started after a short introductory video from the experimenter. To prevent children's drop of attention due to tiredness, we also introduced a pause screen after the 5 first test trials, where they could stop as long as they wished. There was also a 'leave' button that would stop the study anytime pressed.

The experiment lasted for 20-30 minutes (depending greatly on the duration of the pause). All sessions were video-taped using participants webcams, and these videos were sent and kept in a secure server from Universitat Pompeu Fabra which was set up for this purpose.

You can see an example of the experimental material in the Supplementary Material folder.

Participants were included in the sample only if they had completed at least one trial per condition (UE-superset, UE-subset, DE-superset, DE-subset, and a control).

4.3.2 Results

Data from 122 children (96 Catalan and 27 Italian) were recorded providing a total of 1193 trials. All participants who made more than 49% errors in the Memory/Conditional Question (18 children) and those who failed to select the control child more than 49% of times (9 additional children) were excluded. We excluded a total of 27 children (19 Catalans and 8 Italians) [3].

Out of the remaining children, individual trials were excluded for parent intervention (41 trials, 2.14%), because children selected filler characters (9 trials, 0.47%), because they did not select the control character (86 trials, 9%), or because they did not correctly answer the Memory/Conditional Question (194 trials, 20.3%).

The correctness of the Memory/Conditional Question depended on the condition of the trial. In DE trials, the answer was considered correct if children repeated the set they had been asked to repeat (for example, if they were told that 'John is eating pizza with olives and mushrooms', the correct answer was 'pizza with olives and mushrooms'), or if they mentioned only the superset ('pizza with olives' or 'pizza' in that case). In UE contexts, a Memory/Conditional Question was considered to be correctly answered if children mentioned the exact response given by the conditional (for example, if told 'if John goes into the the tent, he eats pizza with olives'

[3] We judge these high numbers of exclusion to be related with the online nature of the study, and specifically with the fact that parents decided the time to take the task: indeed, 15 of the excluded children (56%) took the study on a school day later than 6 PM., and it is possible that they were too tired to pay attention.

they responded ‘pizza with olives’) or if they mentioned the superset (‘pizza’); for UE-Superset condition, also a subset answer was accepted (‘pizza with olives and mushrooms’). In UE control trials, an ‘I don’t know’ response was considered to be correct, as well as all responses in which children explicitly excluded the target set (e.g. answering ‘carrots’ or ‘tomatoes’ or ‘pizza pepperoni’). 150 UE trials and 44 DE trials were excluded.

Thus, after filtering, 96 children were analyzed (77 Catalans and 19 Italians): 22 3-year-olds (M= 40.4 months, range= 34-46 months), 23 4-year-olds (M= 53.4 months, range= 48-59 months), 25 5 year-olds (M= 65.1 months, range= 60-71 months), and 26 6 year-olds (M= 76.5 months, range= 73-82 months). The total number of trials analyzed was 658.

To ensure that children were engaged and followed the task, we first inspected responses for the Memory/Conditional Question in DE context trials. In this condition, responses of the Memory/Conditional Question had the function of helping children’s memory, but they also mark their engagement to the task. Of the 97 trials, children answered correctly 78 (80% of the times). Also the correct response rates for the Memory/Conditional Question in the experimental DE trials were very high (subset: 173/193, 89.6%; superset: 172/194, 90.6%). These results suggest that overall children were engaged in the task. Also anecdotical evidence from parents, who mostly commented that their children were motivated, engaged, and in some cases wanted to continue the study when it was finished, signal a high level of engagement.

Then, in order to make sure that our results were not due to a poor understanding of conditionals, we inspected the responses for control trials and for the Memory/Conditional Question for UE trials. Since in Control conditions the antecedent was false, and the Entailment question always asked about the consequent, we expected children to not select the target character in control questions, as adults did (see section 4.2.2). Indeed, selection of the target character in control questions was very low (UE control; M= .11, sd=.32; DE control; M = .23, sd=.42)., so we may assume that children demonstrated some knowledge of conditionals in the Memory/Conditional Question in UE contexts. Out of 92 children who answered the UE control trial, 53 (60%) gave an acceptable response in the Memory/Conditional Question. Twenty of them answered ‘I don’t know’ while the others answered explicitly excluded the target set (i.e. the one that had been mentioned in the conditional sentence). The rate of correct responses for the Memory/Conditional Question for the experimental UE trials was similar (superset: 135/190 (71%) correctly answered trials in Memory/Conditional Question, subset:132/191 (69%) correctly answered trials in Memory/Conditional Question. Children who did not correctly answered such questions were mainly either refusing to answer or randomly naming objects. However, around a tenth of the data, were incorrectly reporting the subset (i.e. the 'pizza with olives and mushrooms' when they had been told about 'pizza with olives'. This fact may indicate that these children were enriching their inferences with extralinguistic assumptions). These results also suggest that children's understanding of the conditionals was not unlike that of adults.

We finally inspected children's responses in the Entailment Question. We analyzed the selections of the target character depending on the context (UE vs DE) and condition (superset vs subset), by taking the mean of the selections for each child (Figure 4.8). Preliminary analyses on the Catalan and Italian participants showed very similar results; for this reason, as well as for the difference in size of the two groups, we collapsed the two groups. In general, visual inspection of the data shows that UE context trials, and especially in UE-superset trials, the target character was selected more than in DE trials (UE-subset: M = .56, sd = .47; UE-superset: M =.85, sd = .33; DE-subset: M= .44, sd = .45; DE- superset: M = .36, sd = .41).

With these data, as we did for adults, we performed a Generalized Linear Mixed Model for binomial data, including selection of target child as dependent variable. We collapsed the data of the 3-4 year olds into one group and of the 5-6 year olds into a separate group, in order to gain more power. The maximal converging model included context (UE vs. DE), Condition (Superset vs. Subset) and Age group (3-4 year olds vs. 5-6 year olds), as Fixed effects as well as their interactions. Participant and Story were included as random intercepts. All fixed effects were coded with contrast coding. We found a significant effect for Context (β = -2.002, se= .27, χ^2(1) = 70.12, p <.001;. M=.70, sd=.43 in UE; M=.0.40, sd=.43 in DE)Condition was also significant (β = -.66, se= .25, χ^2(1) = 7.32, p <.001; M=.50, sd=.46 for subset; M=.59, sd=.45 for superset), as well as the interaction between the Condition and Context (β = 2.3, se= .51, χ^2(1) = 22.54, p <.001). There was no effect of age group.

Inspecting the interaction, pairwise comparisons using Wilcoxon rank sum test with continuity Bonferroni correction showed that UE-superset was different from all other conditions, including UE-subset and DE-superset (all p's <.001), thus confirming that children were sensitive to Condition in UE context. However, unlike adults, there was no difference between Conditions (subset vs superset) in Downward contexts (p=.94). Interestingly, children showed a tendency (p=.07) to

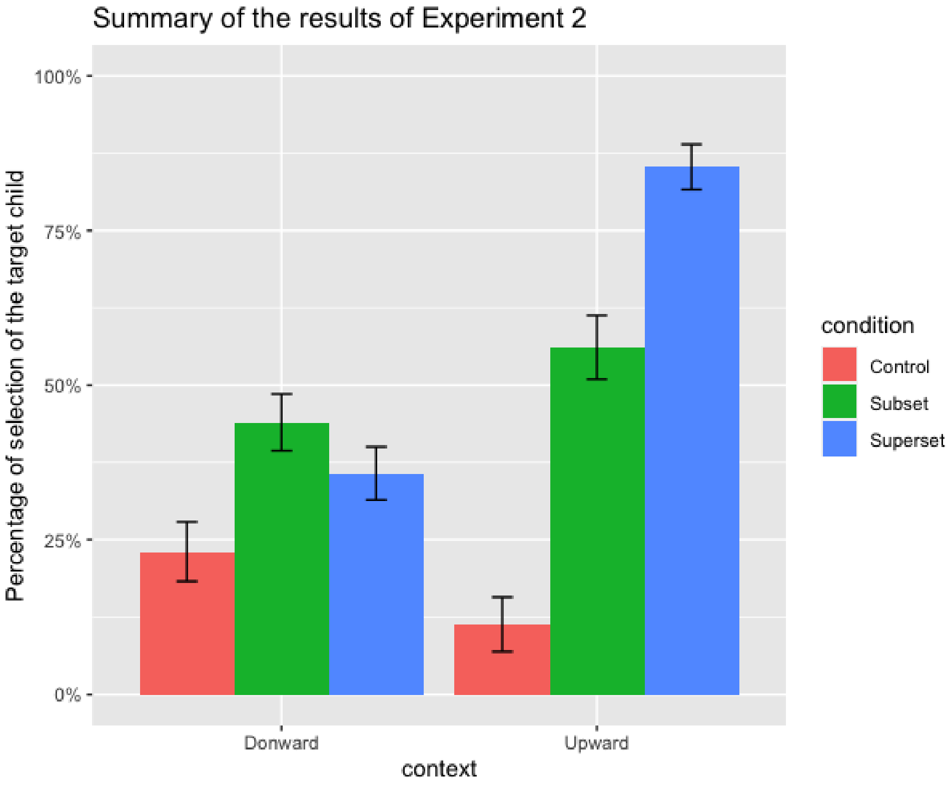


Figure 4.8: Mean of selections to the target subject divided by context and condition. Error bars indicate standard error (se).

select the target character more in UE-subset than in DE-subset, contrarily to our expectations.

Because adults data showed that the effect for DE-subset was dependent on temporal block, we checked possible block effects in children as well. Cochran's Q test revealed no differences between the first and the second part of the study for the DE-subset data (Q(1)=.25, p=.6). Cochran's Q tests for the DE-superset, the UE-superset nor the UE-subset conditions also revealed no differences between the first and the second part of the study (all p's $>$.5).

We then investigated individual patterns of response. We considered the data of all children who had at least one valid response for each of the four experimental conditions (UE-superset, UE-subset, DE-superset and DE-subset; 62 children), taking the mean of each experimental condition. Figure 4.9 shows the individual patterns. We classified children in the same categories as adults (see section 4.2.2), adapted to the fact that they had less responses. Furthermore, we included a new category, 'all-responders' for children who always, or almost always selected the target character. For instance, children from 1 to 7.

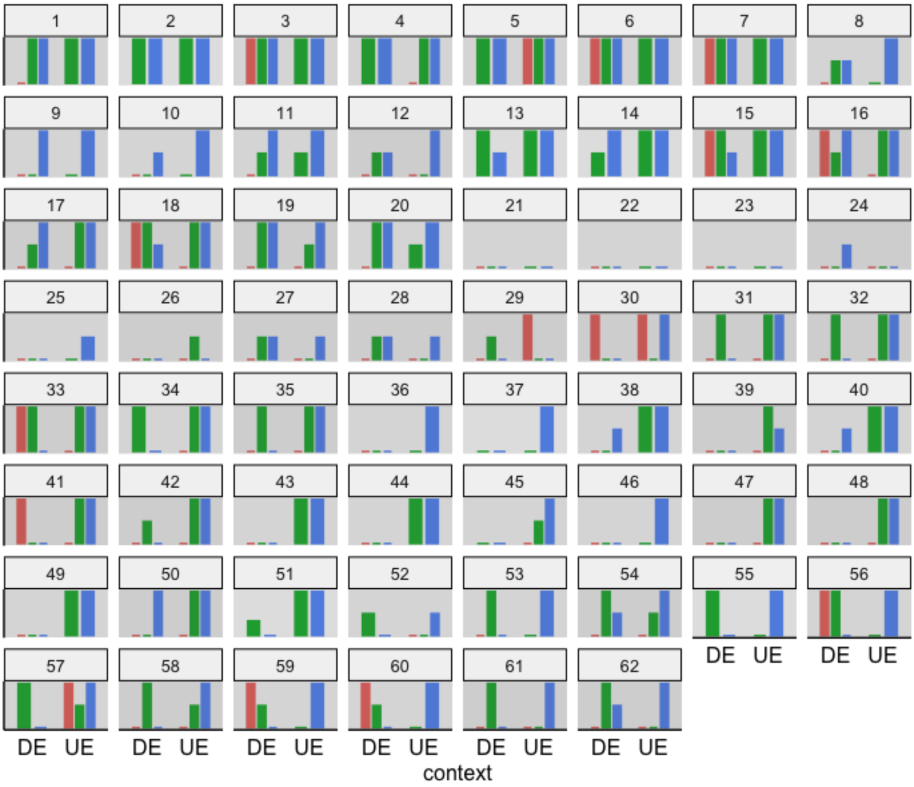


Figure 4.9: Individual patterns of response for each child. Keep in mind that children had 2 trials per pair of condition- context in experimental trials and that we plotted their means. That means that a bar can be at 1 (if the child selected the target character in all her trials for that particular context-condition pair), at 0.5 (if the child selected the target character in one of the two trials for that particular context-condition pair) or at 0 (if the child never selected the target character for that context-condition pair). For control conditions (red bars) only 0 or 1 is possible since our participants had only one trial DE-control and one UE-control. Please note that if the blue and green bars are thick it is because the control conditions for this particular child were taken away from the analysis because of any of the errors described at the beginning of this section. We were mainly interested in the experimental trials, but in order to decide whether a child was an all- or a non-responder we also took the control condition into account.

In order to test whether developmental changes were present across the tested ages, we checked whether the classification differed as a function of age (see Table 4.3).

Response pattern	3-4 year olds	5-6 year olds	Total
Validity responders	3	8	11
Superset responders	4	4	8
Subset responders	2	1	3
Upward responders	6	8	14
All-superset responders	1	4	5
None responders	4	2	6
All responders	6	9	15

Table 4.3: Number of subjects who fell into each response pattern.

A Fisher Exact Test revealed no relationship between the distribution of these patterns and children's age group (p=.65). Therefore, neither in the results nor in the individual patterns of response did children differ by age.

4.3.3 Comparison between adults and children in the interpretation of UE and DE contexts

We finally compared results in Experiments 1 and 2. Figure 4.10 recapitulates both results in a single graph. For this comparison, we ran a Generalized Linear Mixed model, with Context (Upward vs Downward), Condition (Superset vs Subset) and Experiment (Experiment 1, Adults vs Experiment 2, Children), as well as their interactions as fixed factors. All fixed effects were coded with contrast coding. Participant was included as intercept and by random slope by Context and by Condition; and Story was included as random intercept. We found a significant effect for Context (β = -1.39, se= .19, $\chi^2(1)$ = 50.25, p <.001), for Condition (β = -.88, se= .17, $\chi^2(1)$ = 26.9, p <.001) and for Experiment (β = -.69, se= .31, $\chi^2(1)$ = 5.81, p =.016). There was also a double interaction between Context and Condition (β = 4.54, se= .37, $\chi^2(1)$ = 26.9, p <.001), and a triple interaction between Context, Condition and Study (β = 5.05, se= .74, $\chi^2(1)$ = 51.73, p <.001). The double interactions of Context by Study and Condition by Study were only marginal ($\chi^2(1)$ = 3.17, p =.08 and $\chi^2(1)$ = 2.9, p =.09).

Pairwise comparisons using Wilcoxon rank sum test with continuity Bonferroni correction showed that children and adults differed in both invalid conditions (UE-Subset, p<.001, and in DE-Superset, p=.004), whereas in valid inference conditions (UE-Superset and in DE-Subset) there was no difference. Children gave more invalid responses in both conditions (Adults' DE- superset: M = .15, sd = .21; Children's DE- superset: M = .36, sd = .41; Adults' UE-subset: M = .15, sd =

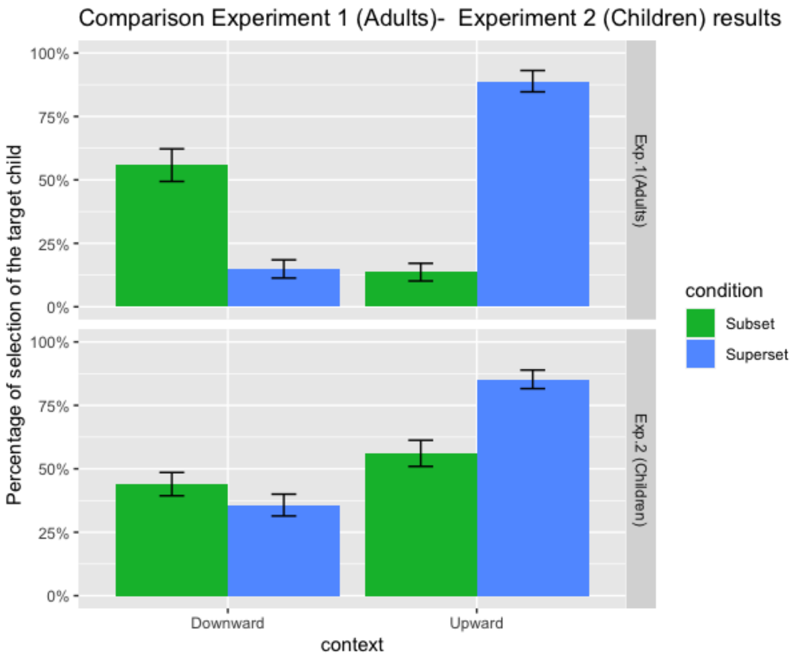


Figure 4.10: Comparison between adults' and children's results.

.22; Children's UE-subset: M = .56, sd = .47).

Finally, we compared the patterns of responses in both adults and children. In order to make a fairer comparison, we excluded from this analysis all participants that had patterns that seem insensitive to our experimental conditions (None responders, 7 children and 1 adult, and All responders, 14 children). With the remaining data we performed a Fisher's Exact Test for Count Data. It revealed that the distributions were different (p =.005).

4.3.4 Discussion

Experiment 2 aimed at testing children's sensitivity to abstract entailment patterns given by upward entailing (UE) and downward entailing (DE) contexts. Our results indicate that children were indeed sensitive to the different entailment patterns that arise in UE vs DE, because they differentiated between contexts in their responses.

This sensitivity, however, is strongly attenuated in DE contexts. In them, children did not modulate their responses, failing to differentiate between valid (subset) and invalid (superset) inferences. This result, somewhat surprising, is strenghtened by the fact that, in individual responses, only around 20% of our participants showed a pattern that can be fully assimilated to the hypothesized pattern (that is, that of accepting the UE-superset and the DE-subset inferences but not the reverse conditions).

These results can also be read from a different angle. Children seemed to strongly prefer superset inferences in Upward contexts, while they showed no differentiation of contexts in subset inferences. Indeed, children who preferred superset inferences in Upward rather than in Downward contexts comprised the 55% of our sample.

It is important to point out that age was not an important factor in the results: there were no noticeable differences whether looking at the individual patterns of responses or in average responses.

Individual patterns revealed a general preference for Upward contexts, a condition in which children also complied with the expected pattern of computing more valid than invalid inferences. This interesting preference, however, could be related with a structural fact: in conditional sentences, like the ones we used in our studies, the Upward context occurs in the consequent of a sentence, while the Downward context is present in the antecedent. Usually, and also in our studies, the consequent follows the antecedent. It is possible that this forced temporal sequence made the implications in UE contexts more salient for children, or that they would facilitate recollection for the last part of the sentence. Albeit a full account in terms of memory degradation is partially controlled by the Memory/Conditional Question, which in DE Contexts ensured memory for the question and improved its retention, we conducted a third study to directly explore the issue of the role of potential temporal effects in processing conditionals. In order to do this, we exploited the fact that languages allow flipping the order of the antecedent and consequent in a conditional without altering its logical status.

4.4 Experiment 3: The role of order of the linguistic material in children's interpretation of entailment

In Experiment 3, we presented the same contexts but we inverted the order of the antecedent and consequent in the conditional sentences eliciting children's response. If saliency or memory factors were responsible for the difference in the acceptance of DE and UE patterns of implication revealed by Experiment 2, then we should find that children better understand DE patterns and reduce performance in UE patterns. Alternatively, if the results of Experiment 2 were due to a different understanding of the nature of the implications in DE and UE contexts, then we should confirm the overall shape of the results obtained in Experiment 2.

4.4.1 Methods

Participants

Thirty-four Catalan children, from 3 to 6 years of age, recruited as in Experiment 2, participated in this Study. Of these, three were excluded because they did not finish at least 5 trials (out of 10), and 4 because they did not meet the minimum criterion for inclusion (see section 4.4.2 for details). The remaining 27 children (5 3-year-olds, 9 4 year-olds, 9 5-year-olds and 4 6-year olds; M = 4.5; 16 females), came from Catalonia, most of them from Barcelona and the great Barcelona area (21). Ten of them were Catalan speakers, while the majority (17) were bilingual Catalan-Spanish (15), and Catalan-French (2).

Procedure and materials

The procedure and materials were identical to those of Experiment 2 (see section 4.3.1). The only change in the experimental material was that the order of the antecedent and the consequent in the conditional sentences was flipped. Thus, sentences like:

(44) If John enters into the tent, he eats pizza with olives. (Upward context)

(45) If John eats pizza with olives he gets stains in his t-shirt. (Downward context)

Were transformed into:

(46) John eats pizza with olives if he enters into the tent. (Upward context)

(47) John gets stains in his t-shirt if eats pizza with olives. (Downward context)

The same speaker who recorded the audios in Experiments 1 and 2 recorded the novel sentences, which replaced the conditionals that we had in Experiment 2. The rest of the material and its presentation was identical to that of Experiment 2 (see section 4.3.1 for further details).

4.4.2 Results

Thirty-one children participated to Experiment 3, for a total of 319 trials. Criteria of exclusion was identical to those of Experiment 2. Accordingly, we excluded 2 participants for excess of errors in the Memory/Conditional Question and 2 who who failed to select the control child as per the criterion therein specified.

The responses of 26 children were retained for analysis: 5 3-year-olds (M= 40.9 months, range= 38-46 months), 9 4-year-olds (M= 50.7 months, range= 48-58 months), 9 5 year-olds (M= 63.2 months, range= 60-65 months), and 4 6 year-olds (M= 79.2 months, range= 74-83 months).

From the data of the analyzed children, 3 trials (.05%) were excluded for parent intervention, 21 trials (7.2%) for failure to select the control character and 50 trials (18.32%) for failure to correctly answer the Memory/Conditional Question. As in Experiment 2 (see section 4.3.2), these criteria led to exclude more UE than DE trials, due to the Memory/Conditional Question (41 exclusions in UE, 9 in DE). This resulted in a total of 205 trials which were retained for analysis.

We ran the same analyses as in Experiment 2. For the Memory/Conditional Question, in DE-control trials children answered correctly 20/27 questions (74%) and for the experimental DE trials 48/54, or 88.89% for the subset and 54/55, or 98.18% for the superset, confirming a high level of engagement. For UE trials, in the control trials, selection of the target character was again very low (UE control; M= .1, sd=.32; DE control; M= .1, sd=.31) and 14 out of 27 (52%) gave an acceptable response (8 'I don't know' answers and 6 answer referring to some member of the targeted set). For the experimental UE trials 41/55 responses (or 74.5%) were correct in the superset and 41/55 (74.5%) in the subset. These results are comparable to those of Experiment 2 (see section 4.3.2), hence suggest the same conclusion about the validity of our results and also that both studies are highly comparable.

The crucial responses were those of the Entailment Question. Figure 4.11 present a summary of the results. Just as in Experiment 2, in UE context trials (especially in UE-superset) the target child was selected more then in DE-context trials (UE-subset: M= .61, sd = .45; UE-superset: M=.81, sd = .35; DE-subset: M= .48, sd = .44; DE- superset: M= .38, sd = .38).

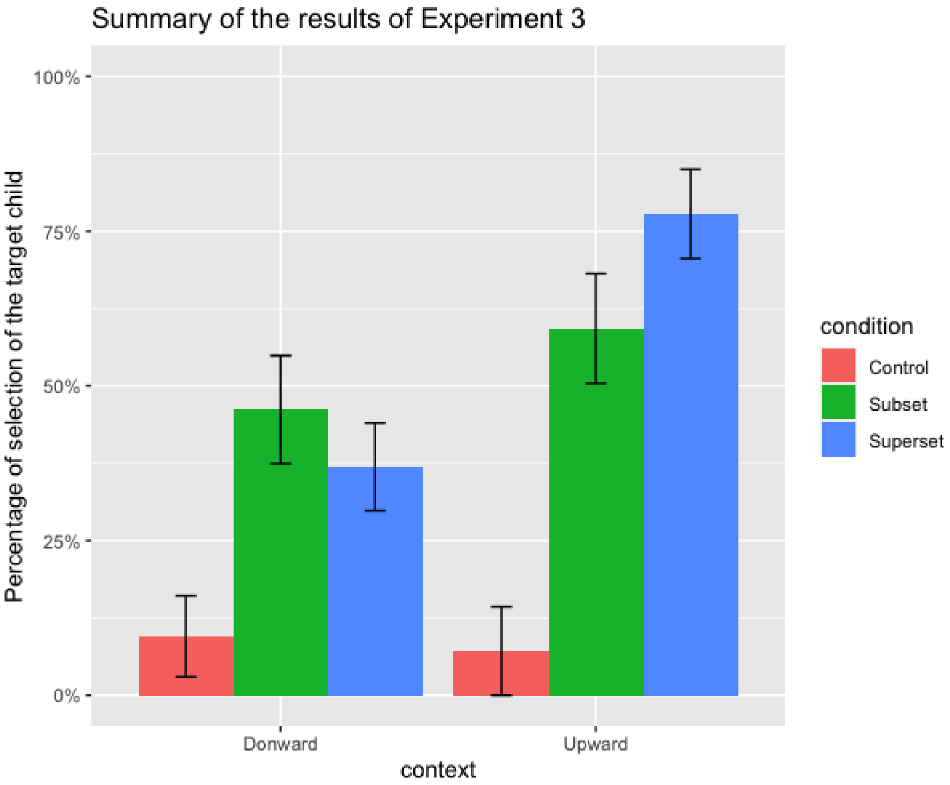


Figure 4.11: Mean of selections to the target subject divided by context and condition. Error bars indicate standard error (se).

A Generalized Linear Mixed Model with the identical design as Experiment 2 gave a maximal converging model identical to that of Experiment 2 and revealed a significant effect of Context (β = –1.72, se= .47, $\chi^2(1)$ = 17.24, p <.001) . a marginal effect of Condition (β = -.72, se= .44, $\chi^2(1)$ = 2.98, p =.08), and no effect of Age. The interaction between Context and Condition was significant (β = 2.3, se= .9, $\chi^2(1)$ = 7.4, p =.006), and a marginal interaction between Context and Age appeared (β = -1.61, se= .9, $\chi^2(1)$ = 3.57, p =.06). Finally, the triple interaction was significant (β = 3.7, se= 1.78, $\chi^2(1)$ = 4.85, p =.03). In short, the results were highly similar to those of Experiment 2, except for the marginal effects of Condition (it was clear in Experiment 2) and of the interaction between Age and Condition (it was absent in Experiment 2).

Pairwise comparisons (Wilcoxon rank sum, Bonferroni corrected) comparing all experimental conditions (UE-superset, UE-subset, DE-superset and DE-subset) separately per age group (3-4 year olds vs 5-6 year olds) showed that younger children did not differentiate between conditions (all p's=1), whereas, for older children, UE-superset was different from DE conditions (all p's <.001, but not from UE-subset (p=.3). A summary of the results by age can be seen in Fig. 4.12.

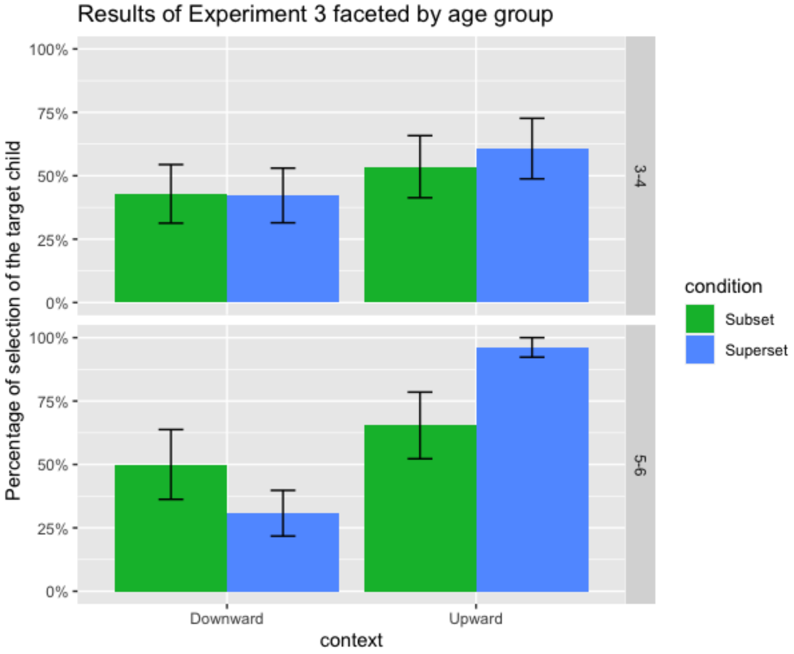


Figure 4.12: Comparison between younger and older children's results.

Although the difference in power between the two studies is considerable, we compared the results of each experimental condition of Experiment 3 with Experiment 2. No differences emerged (pairwise comparison of proportions, all p's>.5).

Individual patterns of response (22 children, Figure 4.13, table 4.4), classified with the same criteria used in Experiment 2, revealed a trend for age differences (p=.096, Fisher Exact test) coherent with the general trend of younger children not differentiating experimental conditions. However, because the trend is non significant, considering the small sample size of each age, we collapsed age groups and we compared the distribution of individual patterns of Studies 2 and 3. There was no difference between the distributions (p=.73, Fisher Exact test), suggesting that neither in the group results nor in the individual patterns of response did Experiment 3 largely differ from Experiment 2.

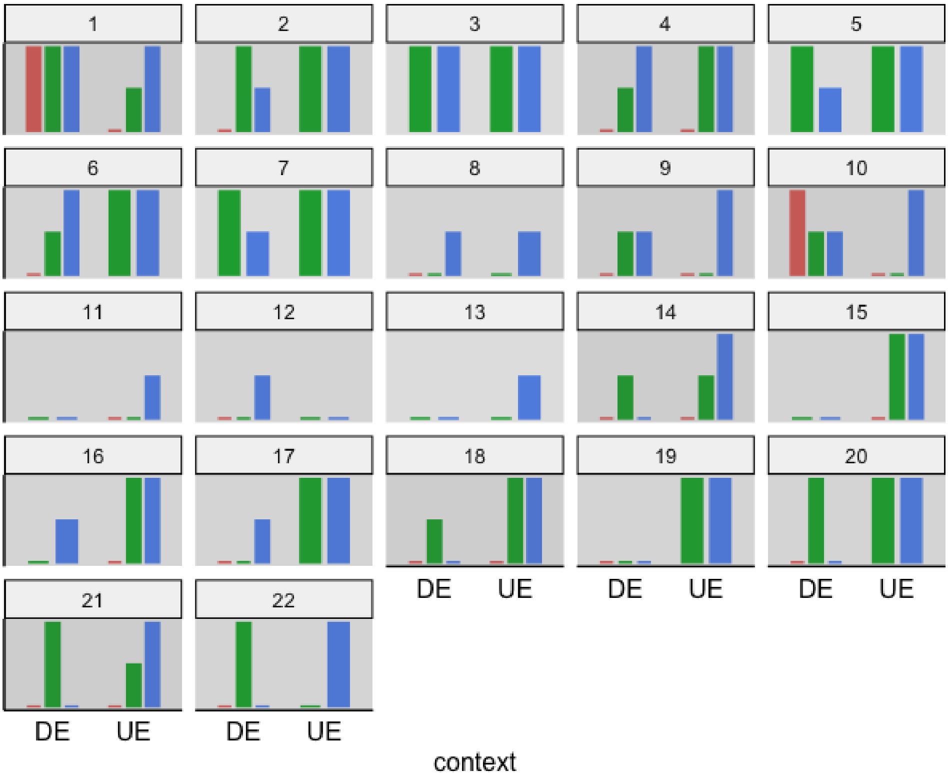


Figure 4.13: Individual patterns of response for each child.

FInally, an important difference between the two studies, concerning age group differences, can be gathered from the excluded trials due to a wrong response in the Memory/conditional or due to failures to select the control character. <ore trials of younger children than older children were excluded (35.5% vs. 21.7%, Fisher exact test, p=.01). This suggests that, overall, younger children found the task more difficult.

4.4.3 Discussion

The aim of Experiment 3 was to test whether the results of Experiment 2, and especially the higher acceptance of valid inferences in UE with respect to DE contexts, could be due to the potential higher salience of UE contexts or to memory effects benefiting them given their prominent position in the sentence, giving them an advantage which could lead to drawing inferences more easily in UE rather than DE contexts. Experiment 3 flips the order of the antecedent and the consequent of

Response pattern	3-4 year olds	5-6 year olds	Total
Validity responders	0	2	2
Superset responders	1	0	1
Subset responders	0	0	0
Upward responders	1	5	6
All-superset responders	1	2	3
None responders	3	0	3
All responders	4	3	7

Table 4.4: Number of subjects who fell into each response pattern.

the conditional sentences, so if factors led to the order, rather than to the semantic interpretation of such contexts, were responsible for this effect, also the tendency to accept valid inferences in the two contexts would flip.

Even though its results should be interpreted with caution because of its small sample size, Experiment 3 indicates that there are no substantial differences induced by the order of the components of a conditional sentence in children's disposition of drawing inferences from them. This appeared both in group results as well as in individual patterns. As in Experiment 2, children tended to draw valid inferences more in UE context than in DE contexts. At the same time, both studies do show that children process conditionals with a particular interpretation in mind, and treat them as carrier of particular patterns of implication. Interestingly, in Experiment 3 a potential difference with Study 2 emerged in the role of age: age group interacted with the disposition of drawing valid inferences in the two studied contexts, it marginally affected also individual pattern results and appeared in the higher exclusion rate of trials and participants among younger children with respect to 5-6 years old. Further, more powerful, experiments are needed to validate this trend, but there are indications that younger children find the inversion of antecedents and consequences in the conditional sentence more challenging.

Overall, however, the general picture emerging from Experiment 3 is that the preference for drawing valid inferences in UE contexts with respect to DE contexts seems to be related with the interpretation that children make of these sentences, and has to be treated as a semantic effect requiring explanation.

4.5 General Discussion

In this set of studies, we aimed at investigating children's understanding of entailment relations. Entailment is very important in linguistic theories (see, for example, Horn, 1989; Chierchia, 2013) because it has been shown to be the unifying glue of several disparate phenomena, such as the interpretation of words with logical valence, the distribution of negative and positive polarity items, or the non-interrogative use of wh-words in languages such as Mandarin Chinese (Su et al., 2012). Entailment has also been claimed to facilitate the acquisition of numerals (Feiman et al., 2019).

That entailment patterns are spontaneously clear – that if somebody eats pizza with pepperoni she is eating pizza – has been assumed to be unproblematic, a psychologically obvious fact immediately accessible by intuition and hence, presumably, available early. However, despite their importance in linguistic theories, and in general, for any theory of thought, the psychological reality of the perception of entailment relations has only been tested indirectly, in studies of children's understanding of words with logical valence (Chierchia et al., 2001; Gualmini, 2014; Su et al., 2012; Panizza et al., 2013), or of the licensing conditions and interpretation of negative polarity items (O'Leary & Crain, 1994). Our aim was, therefore, to directly test whether children are sensitive to abstract entailment patterns and in particular, given their centrality, we focused on whether they would be willing to draw inferences from a set to the superset in UE contexts and from a set to the subset in DE contexts. We operationalized this question by selecting a fundamental form in which such patterns can be manipulated: their appearance in the antecedent or the consequence of conditional expressions.

We first tested our material by studying adults, as a way to establish a baseline for our investigation on younger ages. We verified that with the material we created adults easily draw valid inferences to the superset in UE contexts, and to the subset in DE contexts, and refrain to draw invalid inferences to the subset in UE contexts and to the superset in DE contexts. However, we also found that valid inferences in DE contexts (from the set to the subset) are not as automatic as valid inferences in UE contexts. Indeed, the trend to draw valid inferences in DE contexts augmented as time and experience with the task augmented. This seems to suggest that either our task was not immediately clear to adults, or else that had to overcome a first reading that did not exclusively depend on the logical entailment context. Because in UE contexts valid inference appeared strongly from the beginning of the study, we would tend to favor the first interpretation. A possible explanation of this phenomenon could be that inferences to the subset are inherently more difficult than inferences to the superset. This point has been argued by several authors (Geurts, 2003; Geurts & van Der Slik, 2005), and it has even been related to general, non-linguistic facts, perhaps present even in nonverbal animals (McGonigle & Chalmers, 1996). To our knowledge, no direct evidence for this claim has been obtained with very simple and controlled sentences such as those provided in the current studies. However, again, even assuming that this asymmetry comes from a general bias towards resisting inferences to the subset, the fact that a simple presentation of a few sentences across 40 minutes of one study suffices to overcome it is a sign that it must not be so strong a bias after all.

At the same time, small differences for adults can be complex problems for children across development. Indeed, we found the same asymmetry in the treatment of valid inferences in UE and DE contexts also found in children, across two experiments (see section 4.3.2). As adults, children computed the valid UE inference effectively, but they did not to do the same in DE context,. In it, valid (subset) and invalid (superset) inferences were accepted at an equivalent rate. In our data, these results were stable across the tested age span, although Experiment 3 may hint at a possible developmental trend (4.4.2).

Comparing children with adults we tentatively conclude that the two populations differ less in the acceptance of valid inferences than in the acceptance of invalid inferences. This is especially true for the UE-subset condition, which children tended to accept at higher rates than the valid subset inference (DE-subset). All these consideration lead us to raise three main questions:

1. Why do children seem to prefer drawing inferences in UE contexts?
2. Why do children accept invalid inferences at higher rates than a semantic theory (perhaps any semantic theory) would predict?
3. Why do adults and children draw so few valid inference in DE (DE-subset) contexts?

We will discuss potential answers to those questions below.

4.5.1 (1) Why do children draw more inferences in UE contexts?

In Experiments 2 and 3, children preferred drawing more inferences in UE contexts, for both superset (valid) and subset (invalid) conditions, than in DE contexts.

This preference is also confirmed by individual response patterns: Upward responders were the second more represented group, with around 23% of the children falling into this category, characterized by drawing all inferences in UE contexts but none in DE contexts. Importantly, even in Experiment 3, where the order of the antecedent and consequent of the conditional sentences were reverted, so that DE contexts were listened to last and thus became more salient, children's tendency to preferentially draw UE inferences did not change. It is also worth noticing that the reverse pattern of responses (that is, profiles showing a strong preference for drawing all inferences in DE contexts but none in UE contexts) was entirely absent both in children and adults.

An explanation of these facts could rely on experimental differences between UE and DE which would disfavour DE conditions. The Conditional/Memory question that our participants were asked was different for each condition: while in UE participants had to remember the whole conditional sentence to give an accurate response, in DE participants had to only remember the antecedent, and thus they could ignore the consequent of the sentence (which was needed to correctly answer the Entailment question). This combination of factors could explain why children's responses for DE inferences were around 50%, for both the valid (subset) and the invalid (superset) inference. However, this explanation would also predict that children's results would be around 50% also in the DE control condition, and this was clearly not the case, neither in experiment 2 nor in experiment 3.

Our evidence points at an effect tied to structural constrains on the interpretation of conditionals, so we currently favor the hypothesis that the apparent preference for UE may result form two effects that we describe below.

4.5.2 (2) Why do children accept invalid inferences at higher rates than adults?

The main difference between children and adults was how the two groups accepted invalid inferences (DE-superset and UE-subset conditions). Children seemed to be willing to accept (49) and (51(as a valid conclusion from (48) and (50), respectively more than adults:

(48) a. UE: If Mary hides inside the tent she eats pizza with olives.
b. Mary hid inside the tent.

(49) Mary is eating pizza with olives and mushrooms. (subset)

(50) a. DE: If John eats pizza with olives he gets stains in his t-shirt.
b. John ate pizza. (superset)

(51) John got stains in his t-shirt.

Notice that these inferences, are invalid, but not *contradictory*. Under certain assumptions, they could describe a correct state of affair. Thus for instance, given example (48), if Mary were indeed eating pizza with olives and mushrooms (as in (49)), then the sentence 'If Mary hides inside the tent she eats pizza with olives' would still be true. In the same way, given (50), if John got stains with any type of pizza he ate, the sentence 'If John eats pizza with olives he gets stains in his t-shirt' would again still be true. It is possible that children integrate more extraneous information in their interpretation of the sentences, adding extra step in a potential derivation – something that adults may be less prone to do.

Along these lines, it is also possible that children in our experiments interpreted these sentences in a richer way than what we expected them to do. Specifically, they could have assumed that referring to a set is meant to cover reference to its subset as well. That is, they could pragmatically assume an interpretation that often exists in everyday life, although we tried to avoid in the experiment. To wit, in UE contexts like (48), they could think that the set 'pizza with olives' is used as a synonym for 'pizza with olives and mushooms', especially because in the Entailment Question participants are asked for the subset ('pizza with olives and mushrooms') but never for the starting 'set' 'pizza with olives'. Anecdotical evidence for this account in errors in the Memory/Conditional question may suggest that both adults and children sometimes adopt this interpretation: when participants were asked to draw the inference ('What is Mary eating?') sometimes they responded directly with the subset ('pizza with olives and mushrooms'). Moreover, experimental evidence suggesting that children generally tend to accommodate pragmatically (Katsos & Bishop, 2011) supports our claim.

In a similar way, in DE contexts like (50) participants could also assume that when they are told that 'John is eating pizza', 'pizza' is used as a shortcut for of the subset of pizza that is relevant ('pizza with olives'). Again, some anecdotical evidence in errors in the Memory/Conditional Question for both adults and children seems to support this hypothesis: participants asked 'What is John eating' answered with the set ('pizza with olives') and not with the superset ('pizza'). Indeed, a similar explanation has been given for children's failures in Piaget's class-inclusion task (Politzer, 2016).

Such pragmatic assumption could also explain the high number of children who were ‘all responders’, because under that assumption one would be bound to accept both valid and invalid inferences.

It is not unreasonable to think that a combination of both strategies, deployed by different children, could be at the basis of our results. However, the pragmatic account alone could not explain the whole set of our results, because it would fail to explain their selections in valid inferences, and especially, the weak DE-subset inferences for both adults and children. Thus, even admitting pragmatic factors influencing children’s responses, errors notwithstanding, we are still led to the conclusions that semantic factors are at the basis of the interpretation of UE and DE contexts embedded inside conditionals.

4.5.3 (3) Why is selection of valid DE inferences so weak both in adults and children?

Both in children and in adults, selection of the target character in DE-subset condition was close to 50% and weaker than in UE-superset condition. What is more, in DE contexts children made as many choices of the target character in valid (subset) than in invalid (superset) inference conditions. Individual patterns of choices, in both adults and children confirmed this tendency.

Interestingly, just as there were no participants with patterns of preference for DE (but there were for UE, see section 4.5.1), we did not find any pattern showing preference to respond to subsets only (an ‘all-subset’ mirror pattern of the ‘all-superset’ pattern, see sections 4.2.2 and 4.3.2). Even participants who were ‘subset responders’ in both adults and children were very few (2 adults and 3 children). Compare this number to the 8 adults and 8 children who were ‘superset responders’).

Commonly, DE operators are defined as ‘reversing the canonical entailments’, such that if *x* entails *y*, *Fy* will entail *Fx* whenever *F* is DE (Horn, 1989; Chierchia et al., 2001). This definition assumes that the canonical entailment pattern goes from the set to its superset, that is, the pattern of UE contexts. If the definition captures something of the psychology of entailment, it would be reasonable to expect that DE is the marked case and it is less automatic than the ‘canonical entailment pattern’. However, this begs the question: precisely, one has to explain why inferences to the superset are the canonical pattern.

We observe a difficulty on reasoning about the subset in DE contexts, but not as an unsurmountable difficulty: adults in experiment 1 improved across the blocks of the experiment.

This difficulty could be due to the way we construed the experimental sets, which were compositional (<pizza, pizza with olives, pizza with olives and mushrooms>) rather than lexical (<animal, dog, Dalmatian>): in the construction of the compositional construct 'pizza with olives' the superset 'pizza' needs to be accessed; however to compositionally construct 'pizza with olives' the subset 'pizza with olives and mushrooms' does not need to be accessed. That could lead to different patterns in participants' responses, since the superset would be accessed by default, but not the subset. Such differences should disappear if we tested them with lexical sets, which encode the relationships of set-superset in terms of meaning.

Another possible explanation for the asymmetry between UE-superset and DE-subset entailments has been mentioned above: DE contexts could be intrinsically more difficult than UE contexts for a general, non-linguistic reason (Geurts & van Der Slik, 2005). According to this hypothesis, the crucial difference between the two contexts is that DE contexts force the thinker to focus at the structure of the subsets, and realize that the iteration of this operation cannot go on infinitely but it stop at the grounding point of the empty set. By contrast, one can always climb the ladder of supersets. Stated otherwise, concepts are often organized starting from a natural end point, and a reasoning often consists in measuring the distance away from that point. Thus 'big' stands in opposition to 'small', but 'small' cannot be smaller than non-existence – hence the in the umarked case one asks 'how big' an object is, that is, how distant the measure is from the zero point: 'In many though not all cases, the dimension to which an opposition pair applies has a natural end point, and if it has the direction away from it is the favoured one. Take size, for example. The extent of an object must be greater than zero, or it wouldn't exist, and accordingly the natural direction is from smaller to greater size.' ' UE contexts ask the natural question of the distance from that zero point, whereas DE contexts ask to go towards the zero point, hence its increased conceptual difficulty. While this explanation may be a factor in a learned adult reasoning, it requires an extremely sophisticated understanding of set relations that we consider to be difficult, if not impossible, to attribute to children. At the least, it is a hypothesis which we would be welcome to accept, but which currently has no evidence in its favor.

Looking for a less theoretically loaded explanation, we currently favor another hypothesis. Some evidences coming from other domains, in particular from experiments on visual memory in infants and adults might give some hints at another possible source of the asymmetry between UE-superset and of DE-subset entailments. In these experiments, the computation of the superset is highly automatic, while that of subsets is subject to the limits of visual attention and visual short-term memory (Halberda, Sires, & Feigenson, 2006; Zosh et al., 2011). When presented with an array of different sets of dots and asked to estimate the number of dots in each set, participants do not seem to be able to track a number of subsets higher than two (for 9-month infants, in Zosh et al., 2011) or three (for adults, in Halberda et al., 2006). However, irrespective of the number of subsets, participants always successfully estimated the total number of dots in the screen (the superset). Thus we could hypothesize that, when encountering a hierarchical structure to be encoded in memory, the superset is always the first to be encoded, and then the subsets are, subject to limitations of visual memory. This fact would be supported by investigations on infants' hierarchical visual memory representations (Rosenberg & Feigenson, 2013). The authors found that while infants are able to track up to 3 individual objects in parallel, their binding abilities were 'reduced' at two objects per hierarchical level as a trade off to represent another hierarchical level (two supersets of two sets of two objects).

Under the hypothesis that both phenomena stem from the same operation, our results would be due the encoding of the set, and not a linguistic phenomena. In our design, in the DE-subset condition, we asked participants about a subset of the pizza with olives. However, visually they were presented with a set of food, which included tomatos, pears, carrots, pepperoni pizzas, pizzas with olives and mushrooms and pizzas with olives and ham. If participants visually encoded this set of food in a hierarchical way (food-pizza-pizza with olives- pizza with olives and mushrooms) like studies from Rosenberg and Feigenson (2013) suggest, participants had a number of subsets to encode that overpassed their visual short-term memory abilities. On the contrary, in UE-superset condition participants had to climb up a ladder of the hierarchy 'food-pizza-pizza with olives' which requires (1) to encode less hierarchical levels and (2) to look at the immediate superset (pizza) instead of the immediate subset (pizza with olives and mushrooms), which would be less automatic than looking at the superset.

Encoding the superset by default, as Zosh et al. (2011) suggested, could have the advantage to control all information present in the scene, which, in our task would translate into identifying the domain of reference (e.g. 'we are talking about pizza', for instance). That contrasts with having an exact understanding of the details of each subset which is subject to the limits of visual attention and visual short-term memory, or, in our task, knowing the specific details of the set (e.g. the exact ingredients that pizza has), which would be less important.

4.5.4 Conclusion

The results of this set of studies, in which we investigated whether adults and 3 to 6 year old children could draw inferences based on abstract entailment patterns, suggest that regardless of the differences all groups were sensitive to the entailment context when they computed inferences. However, children of all ages computed a more invalid inferences than adults, a fact that we think can be explained by a combination of pragmatic factors. If our interpretation is right, in essence there is no fundamental difference between the sensitivity of children to entailment, but rather, the transition from a child interpretation to an adult-like lies in the understanding of how to set the proper limits of pragmatic inferences in the contexts we presented to them. What adults can do is to better filter the information and better suspend assumptions which, although made reasonable by the context, are irrelevant to the task of extracting all and only the logical core of the requests. Further research on this hypothesis is required. A second result, even more interesting relates to how adults and children computed valid inferences. While both groups derived a high number of valid inferences in UE contexts (that is, inferences from sets to supersets) in DE contexts (that is, from sets to subsets) both adults and children were less disposed to derive them. We documented this asymmetry by implementing a very simple design in simple grammatical constructions (conditional sentences). Adults improved across blocks, perhaps signaling that the lower degree fo valid inferences in DE contexts does not reflect their stable understanding of entailment in such contexts, but children did not. At the same time, we could exclude that children's poor performance was due to bad memory for the sentences or to the salient position of the consequent in the unmarked order of a conditional. Regardless of the interpretation of the differences between children and adults, the asymmetry between the quantity of valid inferences in UE and DE contexts suggests that valid inferences towards the superset are easily computed than inferences towards the subset. This fact could be related with our choice to use compositional rather than lexical concepts, which could render the superset more transparent than the subset. However, it may also be more general than the interpretation of logical particles in language. It has been suggested that the asymmetry is due to a very sophisticated perception of the lattice organization of the components of a set, down to the identification of its natural end point (Geurts & van Der Slik, 2005). We have suggested a less loaded interpretation, stemming from the observation that the phenomenon bears resemblances with the privileged position of supersets in short-term memory representations (Halberda et al., 2006; Zosh et al., 2011). Although not necessarily alternative, these two accounts are quite different in spirit and consequence. Further research is needed to disentangle them and find their proper role in development.

Thus, while we acknowledge the strong influence of pragmatic factors, in how participants interpret simple conditionals, (section 4.5.2), especially in children, overall our studies show that they coexist with a natural sense of entailment relations, possibly common with with a general sense of the organization of sets and set relations in perception and thought.

Chapter 5

‘Future’ work: Non-linguistic understanding of entailment

5.1 Introduction

In this last chapter, we will present two pilot experiments that turn towards a non-linguistic study of exhaustification. These studies are based on the idea, presented in chapter 1.2.2 that words with logical valence assume an ‘at least’ meaning when they are in a downward entailing (DE) context, whereas they have a more restricted, ‘enriched’ meaning in upward entailing (UE) contexts.

5.1.1 Non-linguistic entailment

As we stated in the previous chapter (chapter 4), our results show that preschoolers were able to understand the different patterns of implications derived by different entailment contexts: participants inferred more towards the superset in UE patterns, thus correctly inferring (53) from (52):

(52) If John hides inside the tent, he eats pizza with olives.

(53) If John hides inside the tent, he eats pizza.

However that effect was much weaker in DE contexts: children did not derive (55) from (54):

(54) If Mary eats pizza with olives, she gets stains.

(55) If Mary eats pizza, she gets stains.

Although interesting, these results cannot tell much about the origins of entailment or exhaustification, because preschoolers have had a lot of experience with language.

To investigate the origins of entailment and exhaustification in general, one needs to resort to infants, who are not yet speaking (see chapter 1.3). Therefore, we need a method to translate entailment relationships to scenes that do not involve language.

To create such materials, we need to translate the entailment context and the set structure (<pizza, pizza with olives>) to non-linguistic components. As we anticipated in chapter 4, conditionals are often related with causal relationships (Menzies & Beebee, 2020), even for young children (Jorgensen & Falmagne, 1992). Therefore, we can rely on causal events to create entailment contexts.

For the set structure, we will rely on quantifiers.

5.1.2 Non-linguistic quantifiers

As we described in chapter 1.2.2, quantifiers, like connectives and numerals, are words whose meaning has a direct logical correspondence.

These words have, along with an 'at least' roughly logical meaning, a second, enriched meaning -which is due to a Scalar Implicature (see chapter 1.2.2 for a more in-deep presentation). These meanings tend to depend on the entailment context in which the word is embedded (see chapter 1.2.2 and also chapter 3 for empirical results): 'at least' meanings seem to appear more in DE contexts (as example (56)), while enriched meanings are more common in UE contexts (like example (57)):

(56) If *some* red balls hit the blue ball, it moves.

(57) If the blue ball hits the red balls, *some* move.

Crucially for our aims, the 'at least' meaning of *some* entails the meaning of *all*, while the enriched meaning does not. In other words, if the rule given is (56) and one sees *all* of the available balls being shot to the blue ball, one will expect that blue ball to move. Conversely, if the rule given is (57), you would probably not expect the blue ball to be able to move *all* balls.

Interestingly, there exist preliminary evidence suggesting that infants can discriminate between ‘some’ and ‘all’ at 12 months: Bonatti and Téglás (2009) used a paradigm in which they habituated 12-month infants to seeing a character opening and closing *some* or *all* of the three doors of the scene (depending on the condition). They then tested with the opposite operator (so if they had seen some doors being opened, they tested infants with all doors being open; and vice versa). Infants looked longer at the opposite operator, which was not related to counting the number of opened doors, because the authors introduced an extra door in the test phase to make sure that their results were not due to infants controlling the number of opened doors.

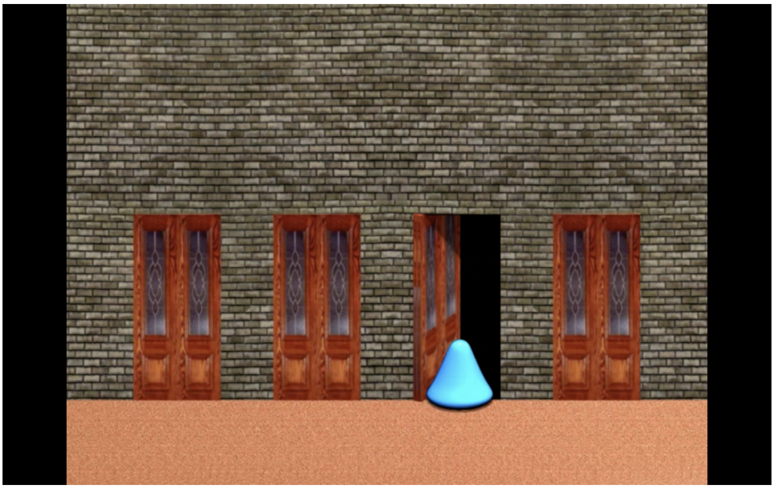


Figure 5.1: Frame from Bonatti and Téglás (2009)’s studies. They habituated 12-month infants to seeing a character opening and closing *some* or *all* of the three doors of the scene, and afterward they tested the infant with the opposite operator. 12-month infants were able to dishabituate in the passage both from ‘some’ to ‘all’ conditions and viceversa. Image reproduced with the permission of the authors.

However, Bonatti and Téglás (2009)’s experiment was difficult to run since each trial was very long. That translated into a high dropping rate of participants, who were losing interest in the scenes before finishing the familiarisation trials. Because of that reason, we decided to depart our attempts to test infants’ understanding of entailment from the same idea but different experimental materials.

5.1.3 The idea behind these studies

We translated Bonatti and Téglás (2009)'s sequential 'some', created by the character by opening doors, to a parallel 'some' conveyed by a subset of balls moving, to make videos shorter. We also introduced a causal event, a hit between our subset of balls and a bigger one, to create a conditional structure that could support the UE and DE contexts. Therefore we created two videos, whose linguistic translation would be example (56) and example (57), repeated here as examples (58) and (59):

(58) DE context: If *some* red balls hit the blue ball, it moves.

(59) UE context: If the blue ball hits the red balls, *some* move.

You can see an example of a video encoding example (58) in figure 5.2a and a video encoding example (59) in figure 5.2c. Moreover, we created also videos encoding example (60) (figure 5.2b) and example (61) (figure 5.2d):

(60) DE context: If *all* red balls hit the blue ball, it moves.

(61) UE context: If the blue ball hits the red balls, *all* move.

This paradigm leads to an asymmetry: *all* and *some* in DE contexts can be equivalent because *some* entails *all*, but the same which does not happen in UE contexts, in which *some* is interpreted as enriched and thus does not entail *all*. Based on that asymmetry we would expect that UE videos appear more distinct than DE videos. Therefore, seeing a UE-all video after a set of UE-some videos should be more surprising than seeing a DE-all video after a DE-some video.

Taking into account Bonatti and Téglás (2009)'s results, it does not seem unreasonable to think that this effect could appear in very young infants (12-month-olds). However, before testing this prediction with infants, we need to make sure that the experimental material we are using is conceptualized in a 'logical' way.

5.1.4 Aim of the pilot studies

The following experiments constitute pilot studies in which we try out the materials we created in adults: we ask adults to describe them, to find whether our stimuli convey a sense of 'some', 'all' and a conditional structure needed for the infants' stimuli.

5.2 Experiment 1a

5.2.1 Participants

We had 19 Italian speaker participants. They were anonymous students and staff from the University of Padova who volunteered to participate.

5.2.2 Methods

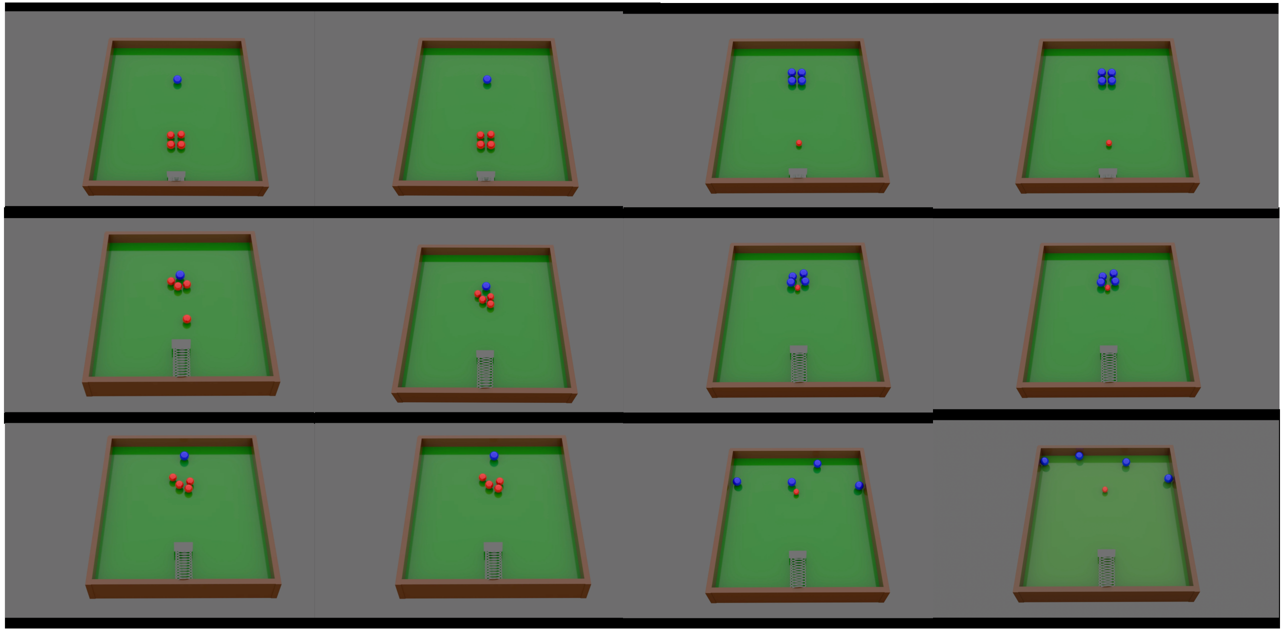
Materials

We prepared four videos using Autodesk Maya in which participants saw a billiard-like table with five balls on (see figure 5.2). One of these balls was bigger and laid apart from the others.

In two of the videos, the big ball was placed at the beginning of the table and the small balls laid at the middle of the table. The big ball approximated the other four and it hit them, making some or all of them move. As anticipated above (see section 5.1.3), these videos tried to pattern the sentence ‘If the big ball hits the small balls, the small balls move’.

In one of these videos (the ‘UE-some’ condition, see figure 5.2c) three of the four balls moved and one stayed still, while in the other one (the ‘UE-all’ condition, see figure 5.2c), all four balls moved.

In the two other videos, the reverse happened. The group of small balls laid at the beginning while the big ball was in the middle of the table. The small balls pushed the big one, making it move. In one of these two videos (named ‘DE-some’), three of the small balls pushed the big one, while the fourth small ball moved slower and did not reach the others in time; while in the other video (named ‘DE-all’), all four balls pushed the big ball. Each of the videos lasted between 5 and 8 seconds. You can see an example of the videos in the the Supplementary Material folder.



(a) DE-some trial **(b)** DE-all trial **(c)** UE-some trial **(d)** UE-all trial

Figure 5.2: Example of test items for DE context (left) and UE context (right). From top to bottom, the first frame of the video of each condition shows a table with five balls on, four red balls near the spring and one blue ball in the middle of the table for DE (figs. 5.2a and 5.2b), and one red ball near the spring and four blue balls in the middle for UE (figs. 5.2c and 5.2d). Afterward, the spring would push the ball(s) and the single ball (UE conditions, figs. 5.2c and 5.2d), some of them (DE-some, fig. 5.2a) or all of them (DE-all condition, fig. 5.2b) would reach the middle ball(s) and push it/them. Finally, the middle ball(s) would move, either the single ball (in DE conditions, figs. 5.2a and 5.2b), some of them (UE-some, fig. 5.2c) or all of them (UE-all condition, fig. 5.2d). This gives four conditions: DE-some trial (fig. 5.2a) which corresponds to 'If *some* red balls hit the blue ball, it moves'; DE-all trial (fig. 5.2b), which corresponds to 'If *all* red balls hit the blue ball, it moves'; UE-some trial (fig. 5.2c) which corresponds to 'If the blue ball hits the red balls, *some* move'; and UE-all trial (fig. 5.2d), which corresponds to 'If the blue ball hits the red balls, *all* move'.'

Procedure

We used the PCIbex platform (Zehr & Schwarz, 2018) to create an online questionnaire, and we sent the link to our participants, who could complete it through any device when the time suited them.

Participants first saw a 'some' video and were asked to freely (and briefly) describe its events. After that, they saw the 'all' video and, again, had to briefly describe it. Finally, we asked participants if they considered that the two videos they had just seen showed the same events and why.

The order of conditions was counterbalanced, with half of the participants seeing UE videos in the first place, half of them seeing first DE videos. The experiment lasted around 10 minutes.

5.2.3 Results

We will first analyze the descriptions of our participants and then their responses to the 'yes-no' question.

The descriptions of the videos mostly focused on the events we wanted (the balls moving and clashing). We asked for four descriptions per participant, which results in a total of 76 descriptions. From these, we took away two of them (from different participants) because they were either blank (one in the UE-all condition) or were generic ('it shows a billiard game', in the UE-some condition). Therefore, we analyzed a total of 74 descriptions, 18 for each UE condition and 19 for each DE condition.

Our participants, in all of their descriptions, mentioned the movement of the group of smaller balls and of the big ball. We focussed on the words our participants used to describe the small balls, which were the ones that were critically changing behavior between the 'some' and the 'all' videos.

We found that 71.62% of the responses (28/36 trials, 14/19 participants in UE; 26/38 trials, 13/19 participants in DE) described the group of small balls by relying on numerals ('3 of the 4 balls', for instance). Quantifiers were used in 9.46% of the trials (3/36 trials in UE, 4/38 trials in DE), by eight different subjects, who did not use them systematically (that is, those using quantifiers in UE-some condition were not the same as who used them for UE-all or DE-some conditions and so on).

Finally, 17.57% of the trials (7/36 trials in UE, 6/38 trials in DE) described the small balls by using non-quantity expressions (such as using the definite article:'the balls' or bare plurals 'balls'). Participants' use of numerals, quantifiers, and non-quantity expressions was similar across the four conditions (McNemar test, p=1).

Interestingly, in the 'some' videos for both context conditions (UE and DE), around half of our participants (10/18 in UE and 9/19 in DE) ignored that one of the smaller balls did not participate in the action (either because it did not move in the UE or because it arrived late in DE). That meant that their description of the 'some' video was either identical or very similar to their description of the 'all' videos. We do not claim that all participants who did not describe that one of the smaller balls did not participate in the action did not notice that fact. However, it seems so for at least some of the participants, based on their descriptions.

We then focussed on the yes-no question, when participants were asked to say whether the events were the same or not. We had slightly fewer responses for both conditions (16 in UE and 15 in DE, since some of our participants left this question blank), and there was no difference between responses in each context condition (8/16 in UE and 8/15 in DE said they were showing the same events).

Among people who considered that the events in the 'some' and the 'all' video differed (8/16 in UE and 8/15 in DE), the most common justification targeted our intended difference: a distinction of the quantity of the balls that moved (that is, that in one of the videos, three of the balls crashed, whereas in the other one, all four acted; 3/8 in UE, 4/8 in DE, 6 people). However, people also pointed at other differences, such as the final position of the balls (3/8 in UE and 2/8 in DE), or the movement of the big ball (1/8 in UE and 1/8 in DE). Interestingly these differences did not exist; because we had controlled for them.

5.2.4 discussion

This first pilot experiment intended to get people's free description of our materials. We aimed to see if UE and DE videos naturally elicited a different treatment, either by their description or by the direct question on whether people considered they were showing the same events.

This was not the case since we found no differences in the descriptions nor the yes-no question. However, we did see that participants largely resorted to numerals when describing the events in the videos, which made it more difficult for a difference based on quantifiers to appear. Therefore, we decided to try to give them some guidelines on their descriptions to see if we could obtain more relevant information.

5.3 Experiment 1b

In this second version of the task, we decided to restrict participants' descriptions to force them to see the events of the videos following conditional structures with a quantifier embedded. To do so, we prevented participants from using numerals, and we forced them to use a word among a list of logical words.

5.3.1 Participants

We recruited 16 Italian speaker participants. As in the previous experiment, they were anonymous students and staff from the University of Padova who volunteered to participate and who had not participated to Experiment 1a.

5.3.2 Methods

Materials

The materials we used in this experiment were the same as in experiment 1a.

Procedure

As in the previous case, we created an online questionnaire through PCIbex (Zehr & Schwarz, 2018) and sent the link to our participants. As in experiment 1a, participants first saw a 'some' video and were asked to briefly describe it ('some' condition). The difference was that in that case, we asked them not to use numerals and instead use one or more words among 'if' ('se'), 'some' ('alcun/a/e'), 'all' ('tutte'), 'no' ('non') and 'but' ('tranne' as in 'all but one'). After the 'some' video, participants had to describe the 'all' video, following the same instructions ('all' condition).

Afterward, unlike experiment 1a, we presented participants the ‘some’ and the ‘all’ videos they had already seen, side by side and asked them to briefly describe BOTH videos in a joint description (’both’ condition), following the same restrictions they had in the first part (that is, not using numerals and using the list of words). We added this task because we hypothesized that participants would need more words or would find it more harder to describe both videos in a single sentence (and hence would have to use more compound sentences) if they had perceived them as more different. Finally, to confirm such indirect measure, participants were asked to explicitly evaluate, on a Likert scale from 1 to 5 (in which 5 was the highest value of similitude), how similar they considered the events of the video were.

As in the previous experiment, the order of conditions was counterbalanced, with half of the participants seeing first the UE videos and the other half seeing DE videos first. The average duration of the task was around 15 minutes.

5.3.3 Results

As in the previous experiment, we will first report our participants’ descriptions and then, their similitude evaluation. Like in experiment 1a (see section 5.2.3), we excluded one description (in DE-some) because it was not describing the event but the situation (’there are some balls on the table, all small except for a bigger one’). Therefore, we analyzed a total of 95 descriptions: 16 for most conditions (UE-some, UE-all, UE-both, DE-all, and DE-both) but 15 for DE-some. We will first focus on the individual descriptions and then on the joint descriptions (conditions UE-both and DE-both).

We divided the action of the video into two parts: the movement of the spring pushing the ball/balls towards the ball/balls in the center, which we will call the ‘antecedent’; and the hit and subsequent movement of the ball/balls that laid in the center, the ‘consequent’. Using this division, we found a difference between the descriptions in UE and DE: most of the descriptions in DE (21/31) did not encode the consequent (i.e. movement of the blue ball due to the hitting of the red), while only 9/32 ignored the movement of the red balls due to the hitting of the blue in UE. A Fisher exact test confirmed that these two proportions were statistically different (p=.002). This fact may indicate that participants only paid attention to the red group of balls and not to the complete scene.

A vast majority of our participants used quantifiers within the list to describe the videos: only three trials (all of them in the UE-all condition) did not include any quantifier. As expected, the quantifiers they used differed: participants used ‘all’ to describe mainly ‘all’ videos (26/32 ‘all’ video trials were described using ‘all’ vs. 4/31 ‘some’ trials). In contrast, ‘some’ was used more in ‘some’ videos (7/31 in ‘some’ videos vs. 3/32 in ‘all’ videos). ‘Some’ videos were also described using ‘not all’ and ‘but’ (19/31 in ‘some’ videos vs. 0/32 in ‘all’ videos). A Fisher test indicated that the distribution was different ($p<.001$).

Having inspected the differences between quantifier use in the descriptions of the ‘some’ and ‘all’ videos, we decided to focus on the differences between UE and DE by inspecting whether there was any substantial difference between the use of quantifiers to describe each condition. We found that ‘all’ video conditions had very similar quantifier uses: 12/16 in UE-condition and 14/16 in DE-all condition used ‘all’, while the remaining used ‘some’ or did not use any quantifier. However, in ‘some’ videos quantifier uses differed: in UE-some most of the people used ‘not all’ or ‘but’ to describe the scene (12/16), with a testimonial number of people using ‘some’ (2/16). In DE-some, however, the number of people using ‘not all’ and ‘but’ dropped (7/15), and the number of people using ‘some’ increased (5/15). Interestingly, in this context, the use of ‘not all’ and of ‘but’ is equivalent to an enriched meaning or ‘some’ (’some but not all’). Since our theory predicts that the enriched meaning of some is more frequently used in UE, while DE contexts favor an ‘at least’ reading, we checked whether the difference in ‘some’ vs ‘not all’/’but’ use could pattern this distinction. However, such difference was not statistically significant (McNemar test, χ^2 = 1.33, p = .25).

We then turned to the analysis of the second part of the experiment, when participants were asked to describe the ‘some’ and the ‘all’ videos together, and the Likert evaluation of the events’ similarity in both videos.

Similar to what happened for the individual descriptions, our participants described the consequent part of the videos more in UE (11/16) than in DE (5/16; McNemar test, χ^2 = 4,16, p = .04).

We checked whether our participants used a single or compound sentence to describe both videos together. We hypothesized that if participants perceived a larger difference between the ‘some’ and the ‘all’ videos in one context condition, they would need to use compound sentences more often and probably, more words. However, we found no differences in the length of both descriptions (Wilcoxon signed-rank test for paired data, V = 34.5, p=.27), nor in the number of compound sentences used (13/16 in UE and 13/16 in DE).

We did not find any remarkable difference in quantifier use between joint descriptions in UE and DE (Fisher exact test, p=.9): people mostly used a combination of ‘all’ + ‘some’ (5/16 in UE and 6/16 in UE) or ‘all’+‘not all’/‘but’(5/16 in UE, 7/16 in DE).

Finally, there was no difference either in the Likert scale evaluation, between UE and DE (UE, median = 3, mean = 3.4; DE, median= 3, mean = 3.2; Wilcoxon signed rank test for paired data, V = 3, p=1).

5.3.4 Discussion

In this pilot experiment, we wanted to see whether differences between UE and DE arose by guiding people to describe our material with quantifiers. We also modified the categorical yes-no question from experiment 1a to use a more fine-grained scale, to capture more subtle differences between UE and DE.

Unlike experiment 1a, this time our participants mostly resorted to quantifiers when asked to describe the videos, both when describing them individually and jointly. We also saw that participants mostly used ‘all’ when describing ‘all’ videos and used ‘some’ or ‘not all’/‘but’ (which would be equivalent to an enriched ‘some’: ‘some but not all’) when they had to refer to the ‘some’ videos. Since our theory predicts that the enriched meaning of some is more frequently used in UE, while DE contexts favor an ‘at least’ reading, we checked whether the distribution of ‘some’ vs. ‘not all’/‘but’ differed between UE and DE contexts, but it did not seem to be the case. However, even though our participants described the videos with quantifiers, as asked, it could still be the case that they first encoded them with numerals, which was their natural tendency (see section 5.2), and then ‘translate’ that encoding into quantifiers.

We noticed, moreover, that only a few of our participants were encoding the scenes in a ‘conditional’ like structure. Especially in DE contexts, our participants tended to neglect the movement happening due to the hit and only wrote about the group of small balls moving.

We decided to run a second series of experiments to address these issues. We created new videos in which we increased the number of balls beyond the subitizing range (to prevent counting). We also changed our participants’ instructions asking them to use conditional sentences in their descriptions.

5.4 Experiment 2a

In this second series of pilot studies, we modified our experimental material by increasing the number of small balls (so they could not be subitized). Moreover, we created six different versions of the 'some' videos to help our participants to focus on the abstract distinction between 'some' and 'all'. Furthermore, we restricted our participants' descriptions by asking them not to use numbers and to start their sentences with 'If'.

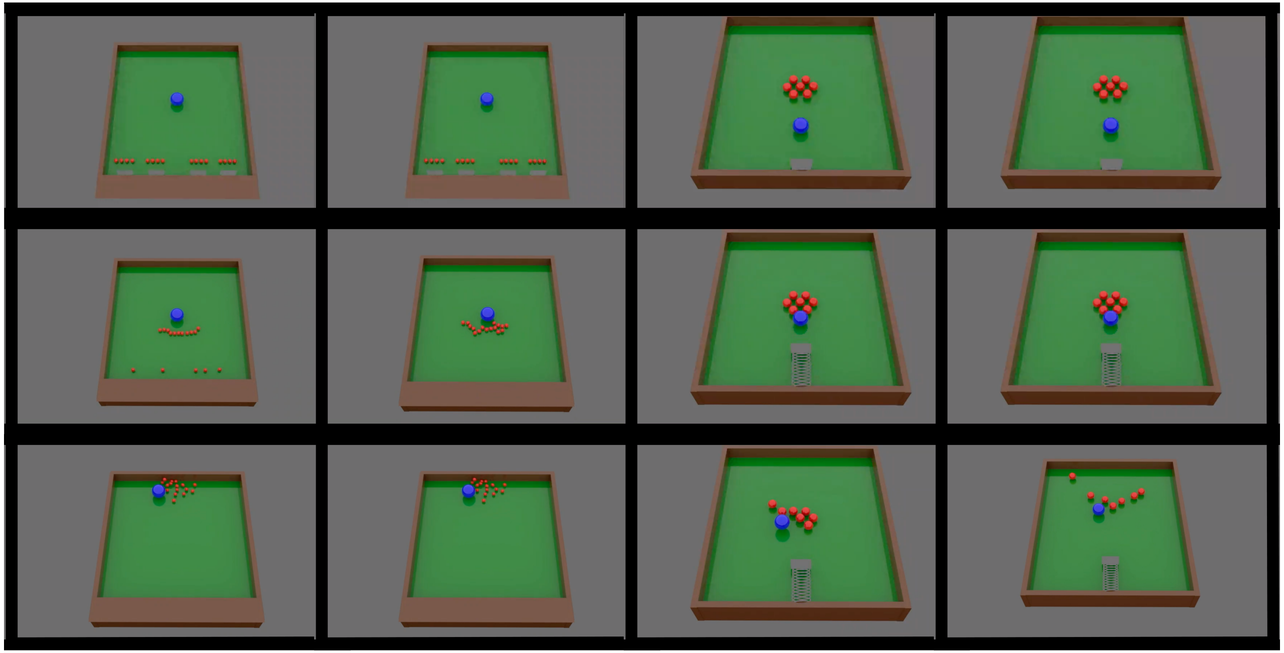
5.4.1 Participants

We had two groups of participants that saw two different versions of the task. The first group was composed of 30 voluntary Italian speaker participants from the University of Venice, but we excluded one because the descriptions he gave were not following the instructions. The resulting 29 (23 females, mean age 28.7, range = 21-54) were proficient Italian speakers. They were asked to answer the questionnaire online but not rewarded. They saw a within-subjects version of the task, in which they evaluated first one context condition and then the other. Moreover, we included a second group of 60 Italian speaker subjects (22 females, mean age = 30.8, range= 19-64) recruited online through Amazon Mechanical Turk. They were paid $5 for their contribution. They saw a between-subjects version: they were only exposed to one context condition (UE or DE).

5.4.2 Methods

Materials

We created new versions of the videos we used in experiments 1a and 1b, using the software Autodesk Maya. To prevent participants from conceptualizing the balls by counting, we increased the number of balls to 16 in DE videos and 7 in UE videos. Each of the videos lasted for 5 seconds. We further created six different versions of the 'some' video, in which different balls were moving and towards different directions, to prevent participants from fixating on details that were not important for the study (direction, speed, number of unmoving balls...). See figure 5.3 for an example of the material. You can see an example of the complete videos in the the Supplementary Material folder.



(a) DE-some trial **(b)** DE-all trial **(c)** UE-some trial **(d)** UE-all trial

Figure 5.3: Example of test items for DE context (left) and UE context (right). From top to bottom, the first frame of the video of each condition shows a table with some balls on, a group of red balls near the spring and one blue ball in the middle of the table for DE (figs. 5.3a and 5.3b), and one blue ball near the spring and a group of red balls in the middle for UE (figs. 5.3c and 5.3d). Afterward, the spring would push the ball(s) and the single one (UE conditions, figs. 5.3c and 5.3d), some of them (DE-some, fig. 5.3a) or all of them (DE-all condition, fig. 5.3b) would reach the middle ball(s) and push it/them. Finally, the middle ball(s) would move, either the single ball (in DE conditions, figs. 5.3a and 5.3b), some of them (UE-some, fig. 5.3c) or all of them (UE-all condition, fig. 5.3d). This gives four conditions: DE-some trial (fig. 5.3a) which corresponds to ‘If *some* red balls hit the blue ball, it moves’; DE-all trial (fig. 5.3b), which corresponds to ‘If *all* red balls hit the blue ball, it moves’; UE-some trial (fig. 5.3c) which corresponds to ‘If the blue ball hits the red balls, *some* move’; and UE-all trial (fig. 5.3d), which corresponds to ‘If the blue ball hits the red balls, *all* move’.’

5.4.3 Procedure

Participants first saw all ‘some’ videos of one condition followed by the ‘all’ video. After that, they were shown one of the ‘some’ videos next to the ‘all’ video and told that the first six videos they had seen were part of a group. Then they were asked, in a yes-no question, whether they thought that the ‘all’ video was part of the group. Immediately after that, they were asked to rate, on a Likert scale with 5 levels (within version) or in a continuous slider with 1000 positions (between version), how similar the ‘all’ video was in respect to the ‘some’ videos group. These were the two non-linguistic questions of the task, that always came at the beginning of the trial.

Following their response, participants saw the ‘some’ videos again and were asked to describe them. As in experiment 1b, they had to avoid numerals in their description and to use words with logical content, like ‘some’ (‘alcun/e’), all (‘tutte’), or ‘but’ (‘tranne’). Moreover, unlike experiments 1, we specifically asked our participants to start their descriptions with ‘if’ (‘se’), to encourage a conditional construction.

After describing the ‘some’ videos, participants saw the ‘all’ video again and had to describe that video following the same instructions they had been given for the ‘some’ videos description. Finally, they were presented again with a random ‘some’ and the ‘all’ video, side by side, and asked to provide a third description following the same directrices as before, but this time summarizing all videos together.

In the within version of the task, participants were presented with DE or UE first, counterbalanced, and then with the other condition. In the between version, they only saw one context condition, either UE or DE.

The average duration of the task was around 15 minutes for the within version and around 10 for the between version.

5.4.4 Results

We will first report the non-linguistic questions and then we will enter into the descriptions.

In the non-linguistic questions, participants were coherent in their answers: trials in which the ‘all’ video was classified as belonging to the ‘some’ videos category were also those who judged the ‘some’ and the ‘all’ videos as more similar in the Likert scale (Fisher exact test, p = .007 in the within version, and p=.001 in the between version).

However, contrarily to our expectations, participants classified the ‘all’ video as being part of the ‘some’ group more in UE context condition (17/29 participants; 58.6% in the within version, and 16/30 in the between version) than in DE context condition (13/29 participants, 44.82% in the within version, and 14/30 in the between version). However, a McNemar test confirmed that responses in the within version were not significantly different (χ^2 = 0.75, p = .38), and a Fisher test proved the same for the between version (p = .6).

After that, we analyzed the similarity scale responses. In the within version, participants answered in a Likert scale from 1 to 5, while in the between version participants had a continuous slider to indicate from 0 to 1000 how similar the ‘some’ and the ‘all’ videos were. Again, participants did not seem to differentiate between the UE condition (within version: mean = 3.14, median = 3; between version: mean = 479, median = 485) and the DE condition (within version : mean = 3.07, median = 3; between version: mean = 535, median = 600; Wilcoxon signed-rank test with continuity correction, V(28) =107.5; p =.62 for the within version and W(59)=385.5, p=.46 for the between version).

We then turned to the analysis of the description questions. Out of 174 descriptions in the within version and 177 descriptions in our between version, we excluded 14 (9 from the within version and 5 from the between version, from 9 different subjects) because they were incomplete. We also excluded, from these analyses, 6 descriptions in the within version and 32 from the between version which participants used counterfactual explanations (like ‘if the blue ball had not hit the red balls, none of them would have moved’) because they were not descriptions in a strict sense but were interpreting further. We also had 29 descriptions in the within version and 16 in the between version failing to comply with the instructions (that is, write a conditional sentence containing words with logical valence). Because these sentences were valid descriptions of the scene, they were included in the following analysis. Therefore, we analyzed 159 descriptions for the within version and 142 descriptions for the between version.

As in experiment 1b (see section 5.3.3), we divided the action that had to be described into two parts: the movement of the spring pushing the ball/balls towards the ball/balls in the center, the ‘antecedent’ of the action, and the hit and movement of the ball/balls that laid in the center, the ‘consequent’ of the action. We found that there was a difference between the descriptions in UE and DE: half of the descriptions in DE (26 out of 54 in the within version and 24/51 in the between version) did not encode the consequent -i.e. the movement of the blue ball due to the hitting of the red, while only 14/53 (within version) and 5/42 (between version) ignored the movement of the red balls due to the hitting of the blue in UE. A Fisher exact test confirmed that these two proportions were statistically different (p=.02 and p=.001). Again, and although not that dramatically as in experiment 1b, that may indicate that participants only paid attention to the red group of balls and not to the complete scene. Because of this result we decided to restrict further analyses to only those trials in which the consequent had been described.

Unsurprisingly, we found that the word 'all' was used to refer almost exclusively to 'all' video conditions (used in 'all' trials 34/36 in the within version and 24/34 in the between version vs. in 1/31 in the within version or 4/30 in the between version, for the 'some' trials). The inverse was also true: 'some' was used to refer almost exclusively to 'some' video conditions (used in 16/31 of the within version and 10/30 of the between version in 'some' trials versus 2/36 in the within version and 6/34 in the between version in 'all' trials). Moreover, while 'all' videos were almost exclusively described using the word 'all', 'some' videos were also described using 'not all' and 'but' (13/31 in within trials, 9/30 in between trials). Indeed, a Fisher test confirmed that these distributions were different (both p's<.001).

We then investigated the differences between the words used to describe UE-some and DE-some. A Fisher exact test for count data revealed that participants were using 'not all' and 'but' more in UE-some than in DE-some in the within version (p=.03). This is an interesting finding because our theory predicts that the enriched meaning of 'some' ('some but not all'), which is equivalent to 'not all' and 'but', is typically derived more in UE than in DE contexts. However, the between version of our task did not replicate this result (p=.4).

Finally, we turned our attention to the description of the 'some' and the 'all' videos together. Again, we saw that participants described the consequent (that is the clash and the subsequent movement of the ball(s) in the middle of the table, see 5.3.3 for more details) more in UE (20/27 in the within version and 17/21 in the between version) than in DE (10/25 in the within version and 16/28 in the between version). As in experiment 1b (see 5.3.3), we checked whether our participants used a single sentence or compound sentences to describe both videos together, hoping that, if participants perceived a bigger difference between the 'some' and the 'all' videos, they would need to use compound sentences more often and probably, more words. However, we found no differences in the length of both descriptions (Wilcoxon signed-rank test for paired data, V = 152, p=.68 for the within version, Wilcoxon signed-rank test for unpaired data, W = 146.5, p=.71 for the between version), nor in the number of compound sentences used (McNemar test, χ^2=0, p=1 for the within version, Fisher exact test p=.72 for the between version).

There was no difference in the logical words participants used in their description (Fisher test, all p's $>$.2), although an important number of them used combinations of 'all' and 'some' in their descriptions (14/30 in the within version but 7/33 in the between version). Interestingly, we found that few descriptions of our participants referred to differences/similitudes related with the theory (23/52 in the within version but only 8/49 in the between version), while a similar number of them (20/53 in the within version, 4/49 in the between version) paid attention mostly in the direction and final location of the balls.

5.4.5 Discussion

In this experiment, we wanted to see whether our materials could elicit conditional-like descriptions that showed the effect of asymmetric entailment according to the logical context. It differed from experiments 1 (see sections **??**) because to prevent the two main issues that appeared there, we introduced further guidance in our participants' description (we made them start the sentence by 'if'), and we increased the number of balls presented in the video in an attempt to make the use of quantifiers more natural. Furthermore, we included two non-linguistic questions before the linguistic ones, in which we tried to get whether our participants could indicate the between context condition differences (UE vs DE) in a non-linguistic measure.

As in experiments 1, non-linguistic questions did not show any difference between conditions. Moreover, we also found that a substantial number of people, even if we tried to force them into a conditional structure, ignored the movement of the big ball when it was not necessary to explain the behavior of the small balls. This phenomenon affects more the DE context condition (in which the big ball moves as a consequence of the smaller balls movement) than the UE context condition (in which the big ball *causes* the small balls to move). This fact may indicate that an important amount of our participants did only focus on what was changing in the action (i.e. the movement of the small balls) rather than the full action.

However, unlike what happened in previous experiments, we could appreciate a difference in the use of quantifiers when used to describe UE-some and DE-some conditions: quantifier expressions that are equivalent to the enriched ‘some’ meaning were used more in UE, while bare ‘some’ was used more in DE. That was true for the within version but not for the between version. A potential explanation of this disparity between experimental versions is that our participants benefited from the comparison between UE and DE context conditions.

To confirm this finding and control for the potential effects of order between tasks, we decided to run a follow-up study, inverting the order of the tasks.

5.5 Experiment 2b

In this experiment, we aimed to replicate the results of study 2a, but changing the order of the questions: we first asked the description questions and then the classification ones. Our aim was to replicate the main finding of experiment 2a, that participants used quantifier expressions equivalent to the enriched meaning of ‘some’ more in UE-some condition than in DE-some condition. Moreover, we were interested in seeing potential effects of the tasks’ ordering: whether first describing could affect participants’ non-linguistic judgments.

5.5.1 Participants

31 Italian speakers (13 females, mean age = 29.4, range= 18-59) recruited through Amazon Turk participated in the experiment. An extra participant was recruited but excluded since we doubted she was an Italian speaker. They were paid $5 in compensation.

5.5.2 Methods

Materials

We used the same materials as in experiment 2a (see section 5.4).

5.5.3 Procedure

We used the same procedure as in the within version of experiment 2a but this time we inverted the tasks' order. Thus participants first saw the 'some' videos and then they were asked to describe them with a conditional sentence that included logical words (such as 'some', 'all', 'but', and 'no'). Then they were presented the 'all' video and they had to provide the description, following the same parameters, and finally, they saw a randomly-picked 'some' video and the 'all' one, and were asked to describe them together, with the same constraints as their first two descriptions. Afterward, participants saw all videos again and were asked the two non-linguistic questions: first whether the 'all' video could be part of the 'some' video group, and finally, to rate the similitude, in a continuous slider, between the 'all' and the 'some' videos.

Participants completed the tasks for one condition (UE or DE) first, and then they were presented the other, counterbalanced. The average duration of the task was around 15 minutes.

5.5.4 Results

For easier comparison, we will first present the results of the non-linguistic questions, and then the description questions.

As in experiment 2a, our participants were coherent within their classification responses: trials in which the 'all' video was classified as being part of the same group as the 'some' videos were also those with a high similarity score in the continuous slider (Fisher Exact Test $p<.001$).

As in the previous study, we found that there was no difference between conditions in deciding whether the 'all' video was part of the 'some' videos (18/31 in UE, 22/31 in DE, with only 6/31 who classified the 'all' video in the 'some' group in DE but not in UE, which would be our expected result; McNemar's Chi-squared test, χ^2 = 1.13, p = .28).

Unsurprisingly, there was no difference either in the scalar similitude question (median UE= 572, median DE = 547, Wilcoxon signed-rank test with continuity correction for paired data: V = 208.5, p=.63).

We then turned to the description questions. Out of the 186 descriptions, we excluded three because they were incomplete and 16 because they used counterfactual constructions. Additionally, 11 more did not have a conditional form but were included in the following analysis because they constituted accurate descriptions of the scene. That left us with 167 descriptions to analyze.

As in experiment 2a, we divided the action that had to be described into two parts: the movement of the spring pushing the ball/balls towards the ball/balls in the center (the 'antecedent') and the hit and resulting movement of the center balls (the 'consequent'). As in the previous experiments, a substantial number of descriptions (15/57 in the DE and 13/56 in the UE) did not encode the consequent. However, unlike experiment 2a, there was no difference between conditions (Fisher exact test, p=.82). We decided to restrict further analyses to only those trials in which the consequent had been described.

As in previous experiments, we found that the word 'all' was used to refer almost exclusively to 'all' video conditions (36/43 in 'all' videos vs. 4/42 in 'some' videos). The inverse was also true: 'some' was used to refer almost exclusively to 'some' videos (17/42 in 'some' videos vs 3/43 in 'all' videos). However, again, while 'all' videos were almost exclusively described with 'all', 'some' videos were also described using 'not all' and 'but' (16/42). A Fisher test confirmed a difference in the distributions ($p<.001$).

To find whether participants differentiated UE vs. DE in their descriptions, we turned our attention to the words they used in UE-some vs. in DE-some. As in the within version of experiment 2a, participants used more 'not all' and 'but' in UE than in DE (11/21 in UE vs. 4/21 in DE; Fisher exact test, p=.03).

Finally, we analyzed the conjunct description, in which participants described the 'some' and the 'all' videos together. We checked whether our participants used a single sentence or compound sentences to describe both videos together, and found no differences in the length of both descriptions (Wilcoxon signed-rank test for paired data, V = 201.5, p=.52), nor in the number of compound sentences used (McNemar test, χ^2=0, p=1).

There was no difference in the logical words participants used in their descriptions (Fisher test, p=.38), although an important number of them used combinations of 'all' and 'some' in their descriptions (18/39). In that case, we found that more descriptions of our participants referred to differences/similitudes related to the theory than in the previous studies (26/39).

5.5.5 Discussion

This pilot experiment replicated the main findings of experiment 2a. We confirmed that our participants described the UE-some videos with words equivalent to the enriched meaning of ‘some’, while they tended to use bare ‘some’ when asked to describe the DE-some videos, in line with our hypothesis.

However, and in line with what we had found in experiment 2a, this differentiation occurred only in the linguistic descriptions, and not in the non-linguistic questions.

5.6 General Discussion

The series of pilot studies presented in this chapter aimed at testing whether our non-linguistic stimuli could encode entailment effects in a similar way conditional sentences do.

The videos showed to our subjects in all the pilot studies depicted a billiard-like scene in which there was either a single ball that would crash with a group of balls in the center or a group of balls that would crash with a single ball in the center of the table. They tried to emulate, through causality (the clash) and the number of balls that would move (all of them or just some) the entailment effects that result from conditional sentences encoding the quantifiers ‘some’ and ‘all’ in UE and DE contexts, as in examples (62) and (63):

(62) DE context: If *some/all* red balls hit the blue ball, it moves.

(63) UE context: If the blue ball hits the red balls, *some/all* move.

In our studies, we asked subjects for non-linguistic measures and linguistic descriptions of our videos. Table 5.1 is a summary of all the experiments and their results.

Our non-linguistic measures aimed to directly test whether, without any prompt of language, our participants would differently treat UE and DE videos. We asked them to decide, by grading in continuous scales, Likert scales, or dichotomously, the similarity between ‘all’ and ‘some’ videos and their completeness as a group, hoping that they would see fewer differences in DE context conditions than in the UE context condition. That was not the case, for all of our non-linguistic measures showed that participants treated both conditions in the same way.

However, we found more interesting results in the semi-guided descriptions of our participants. Overall we saw that our participants were not encoding the conditional structure, especially in DE videos, since it seems that they tended to fixate only the movement of the group of balls and not the complete scene. This influenced our results since they are dependent on the structural relation setting the logical context that conveys the asymmetric entailment (that is, the quantifier appearing in DE or UE) we were looking for. In other words, we think that, across experiments, our participants did not encode a conditional-like structure for DE videos, but they only encoded the first part of them (in which the group of balls moved). This fact could also play a role in the non-results of the non-linguistic part of the experiment, since encoding a conditional-like structure is an indispensable prerequisite to perceive the differences in entailment in UE and DE videos.

When we excluded participants that had not encoded the movement of the solo ball for both UE and DE (see sections 5.4.4 and 5.5.4), we found an interesting difference in the way participants described ‘some’ videos: participants tended to use words that had the same meaning as the enriched some (‘some but not all’) in UE-some video descriptions, whereas in DE-some videos they resorted to using plain ‘some’.

We have to remember, at this point, that participants’ descriptions, even though guided with some restrictions (like forcing them to use logical words), were free. That is, they could (and indeed did) focus on a potentially infinite number of dimensions to describe them. However, they, maybe even unconsciously, chose different words to describe the UE-some and the DE-some videos; and, more importantly, their word choice was coherent with the asymmetrical entailment pattern that the videos wanted to convey. This effect is only appreciable in the within-participant versions of our task, suggesting that some amount of comparison between context conditions is needed for it to appear. This difference between context conditions, albeit very subtle and indirect, is an indicator that participants appreciated a difference in the entailment relationships driven by UE and DE context conditions.

If we found such effect in their descriptions why did not we find anything in the non-linguistic tasks? A potential explanation of this fact is related to the dimensions of complexity participants had to deal with. In our non-linguistic questions, participants were only asked whether the ‘all’ video could be part of the ‘some’ video group, or else how similar the ‘all’ video was to the ‘some’ videos. These questions are very wide and could prompt an answer based on very different dimensions, such as the visual similitudes of the videos. The videos were objectively very similar (all of them feature balls and a billiard table), but they also featured obvious visual differences (in the 'some' videos a subset of balls does not move or displays a different movement, whereas in the 'all' videos all of them do in the same way). It seems very possible that participants resorted to visual differences and we failed to make the logical distinction salient enough for them.

Once we explicitly highlighted the dimension we wanted our participants to work with (logical words), they came up with differences in their descriptions. However, this effect had no consequences for the non-linguistic judgments: as experiment 2b (see section 5.5) shows, if participants were first asked to describe (and so asked to focus on the logical structure) before having to classify/state the differences between 'some' and 'all' videos in a non-linguistic way, they still found no differences between context conditions.

The fact that we only find subtle differences in the linguistic descriptions points out to the conclusion that language makes it easier to convey these logical patterns. However, it would be, to our judgment, a mistake to think that it is *the only* way to make such differences appear or to advocate for a primordial role of language in the encoding of entailment relations based on the data we just presented. Participants' descriptions pointed out that it was difficult for them even in linguistic descriptions to encode conditional-like descriptions of the event, which is a necessary pre-requisite for the distinctive entailment patterns to appear.

Indeed, our data seems to suggest that access to a logical encoding of the scenes requires an amount of generalization and abstraction that our stimuli only partially conveyed.

With these results in hand, we think that the stimuli are still not ready to be used with infants, since they do not seem to be completely conveying the logical structure to adults. However, we plan to improve them by using procedures that highlight the conditional structure, especially marking the movement of the consequent part (that is, that the hit balls move as a result of the clash). An example of this procedure would be to stop the antecedent balls (that is, the balls that are pushed by the spring) when they hit the consequent balls (the ones that are in the middle). In that way, the only movement (and hence the most salient part) on the scene in the consequent part of the videos would be the moving of the hit balls. We plan to continue this project in this direction.

Experiment		Stimuli	Experimental questions		Results
			Linguistic descriptions	Non-linguistic	
1	**a**	Stimuli with few balls (see fig. 5.2)	Coming first. Free description of the 'some' and the 'all' videos	Coming after the linguistic description. Asking whether the 'some' and the 'all' video showed the same events.	Resorting to numerals to describe. No differences between UE and DE.
	b		Coming first. Guided description of the 'some' and 'all' videos and of both videos together.	Coming after the linguistic description. Evaluation of the similitude of the events of the 'some' and the 'all' videos on a Likert scale.	No differences between UE and DE. Not using conditionals in the descriptions.
2	**a**	Stimuli with more balls (see fig. 5.3)	Coming after the non-linguistic questions. Guided description (asking explicitly for conditionals) of the 'some' and 'all' videos and of both videos together.	Coming first. (1)Yes-no question on whether the 'all' video was part of the 'some' video group and (2) evaluation of the similitude of the events of the 'some' and the 'all' videos on a Likert scale (within version) or on a continuous slider (between version)	No differences between UE and DE in the non-linguistic questions. Difference between UE and DE in the descriptions in the within version.
	b		Coming first. Guided description (asking explicitly for conditionals) of the 'some' and 'all' videos and of both videos together.	Coming after the linguistic descriptions. (1)Yes-no question on whether the 'all' video was part of the 'some' video group and (2) evaluation of the similitude of the events of the 'some' and the 'all' videos on a continuous slider.	No differences between UE and DE in the non-linguistic questions. Difference between UE and DE in the descriptions.

Table 5.1: Summary of the experimental procedures and their results.

Chapter 6

General discussion

Cogito, ergo sum;

claims Descartes' most famous quote. The nature of human cognition is a complex issue that has captivated scholars since the dawn of time. It is not difficult to understand why: our intellect has allowed us to transform the world and adapt it to our needs and wills instead of submitting to its rules.

Within the set of capacities conforming human cognition, there is one that excels on its specificity to our species: language. A relationship between language and thought seems undeniable. However, as we stated in the introduction of this thesis, the number of computational devices and the potential causal relationships these two objects share are more debated.

In this thesis, we tried to shed light on the nature of human thought and its link with language. We started this project with two aims in mind: first, we were interested in investigating the logical capacities supporting language, and second, we wanted to use these logical capacities to access the structure of thought, by describing potential candidates to primitives of thought. It is time to take stock and revise where this journey has brought us.

We departed from a well-established theory about the logicality of language, the theory of exhaustification (Chierchia, 2013), which we discussed in chapter **??**. This theory proposes that several interpretational phenomena that are thought to be idiosyncratic and external to language are not but depend on formalizable computations. The main idea behind it is that a sentence is interpreted against a background of ordered alternative sentences. By calculating the amount of information carried by each of these alternatives, exhaustification negates them (if not leading to a contradiction), and obtains the sentence meaning. The theory of exhaustification has great explanatory power since it applies to the duplicity of meanings of words with logical valence (a paradigmatic case study in the grey area between semantics and pragmatics), but also to syntactic phenomena (such as the distribution of 'any').

We identified four main subcomponents of the proposed computations: retrieval of the logical meaning of the word, generation of a set of alternatives, ordering of these alternatives, and closure of the set. We then, in chapter 1.3, revised what is known about infants' abilities and learning related to our four subcomponents.

This theoretical approach helped us to identify two essential components of the linguistic theory that have also a prominent role in the inferential reasoning abilities: finding intersections between the development and meaning of words with logical valence, and the role of entailment in children's abilities. These were our specific targets in this research.

6.1 The development and meaning of words with logical valence

We devoted chapters 2 and 3 to the development and meaning of words with logical valence, as a prompt to understand the origins and behavior of exhaustification in the environment (words with logical valence) for which the theory was born.

We investigated the development and meaning of words with logical valence in chapter 2. We focussed on 'and' and numerals, words with logical valence that have the further advantage of being logically equivalent, and we studied them from a developmental perspective. The choice of 'and' and numerals as case studies was further justified because, while the process of acquisition of numerals has been extensively studied that of 'and' remains poorly known. Therefore, we compared the understanding of numerals using a well-known experimental method (Wynn, 1992), and the logically equivalent conjunction in children learning their first numerals, in an attempt to compare the process of acquisition of numerals to that of 'and'.

The results revealed the existence of a relation between the understanding of numeral words and the connective ‘and’: children who were struggling to infer the enriched ‘exact’ meaning of numerals were also struggling to infer the ‘only A and B and nothing else’ meaning of conjunctions with a number of conjuncts corresponding to the numerals they had difficulty with. These results do not seem to be explainable by simple memory task demands, because of the sizable amount of children who could understand complex conjunctions but yet were still struggling with the corresponding numerals, and the fact that the performance in the ‘and’ task was higher than in the numerals task.

We also found that in both tasks children seemed to systematically make more errors compatible with an ‘at least’ interpretation, suggesting that they struggled to come up with the enriched interpretation of these expressions. The commonalities in development between the understanding of numerals and ‘and’ that we found in chap. 2 are not easily explained on the basis of the process of the understanding of the numerical series nor on children’s attested problems with alternatives (Pagliarini, Bill, et al., 2018; Barner et al., 2018; Gotzner et al., 2020). Thus, the correlation between the understanding of numbers and complex conjuncts, as well as the number of errors compatible with an ‘at least’ interpretation, seem to point at a common origin in the acquisition of the first meaning children assign to both numerals and multiple conjunctions. The exact nature of this common origin remains veiled; however, the process of exhaustification seems an obvious candidate, as it rules the passage between one and the other meaning of a logical expression.

We found more shreds of evidence of the regularity of such operation in chapter 3. In this chapter, we took a step further to investigate its systematicity and relationship with our second research target, the role of entailment.

Chapter 3 presented a study on the interpretation of connectives (’and’ and ‘or ‘) and numerals embedded in Upward and Downward contexts of a conditional sentence. Our interest in this study was in finding whether participants would make systematic use of entailment relations in the interpretation of words with logical valence and, more specifically, in their estimations of informativeness of the alternatives.

To carry out this investigation we embedded the target words (numerals, conjunction, and disjunction) into the antecedent (DE) or the consequent (UE) of conditionals, and then asked the participants to perform a ‘scaled acceptability judgment task ‘, in which read a conditional sentence, and after seeing the actions performed, they were asked to grade, from 1 to 5, and based on the conditional sentence, how well the target action had been performed.

Our first experiment of that chapter confirmed that the entailment context (UE/DE) modulates the understanding of 'or' in the way the theory predicts: the at least interpretation of 'or' is more accepted in downward entailing contexts. Interestingly, as we confirmed in the second experiment of that chapter (see chapter 3.3), the evaluation of the sentences integrates formal estimations of informativeness that give rise to the modulation of the meaning of the logical words in the way the theory predicts, with pragmatic factors such as being partially satisfied with the contents of a false sentence.

The output of the first two experimental chapters points up at a regular computation taking place in similar environments. It seems to affect semantically-related operators in their acquisition first and their meaning afterward in a systematic way and it seems to interact with other principles of cognition, such as partial satisfaction (see chap. 3).

Is exhaustification encoded in the grammar, as Chierchia (2013) proposed? Our datacannot advocate for the presence of a grammatical operator. However, a growing number of studies focusing on syntactic behavior (Chierchia et al., 2001; Panizza, Chierchia, & Clifton Jr, 2009; Chemla & Spector, 2011; Panizza et al., 2013; Hartshorne et al., 2015; Pagliarini, Bill, et al., 2018) or atypical populations (Arosio, Foppolo, Pagliarini, Perugini, & Guasti, 2017; Hochstein, Bale, & Barner, 2018) suggest that this might be indeed the case.

Setting aside the consequences this pervasion of logic into syntax would have in a theoretical level, which are not our topic of discussion, our studies suggest that the operation of exhaustification is playing a generalized role in language understanding, and may extend beyond it -the role it has with the learning of number suggests so.

Exhaustification can be seen as a closure operation that encapsulates the relevant domain to prevent running into the problem of dealing with infinity every time we face a sentence. The results of our studies suggest a pervasive role of logic into language, like the Logicality of Language hypothesis suggests. Within this framework the existence of a logical structure supporting language is not surprising: it would help infants navigate the linguistic space and ease memory and computation demands when acquiring the language.

Indeed, the need for a closure operation is also observable when dealing with our second case-study, that of the computation of entailment relationships.

6.2 The development of entailment relationships

Children's understanding of entailment relationships constituted our second focussed aim and is the topic we addressed in chapters 4 and 5. Entailment relations play a vital role in language theories since some formal semanticists consider entailment a core property of human languages (Chierchia & McConnell-Ginet, 2000; Su et al., 2012) since all languages contain expressions that restrict their context of appearance according to the entailment context (Ladusaw, 1979; Horn, 1989; Su et al., 2012).

In chapter 4 our interest was to see whether children could take entailment patterns into account. Contrary to previous studies, which had focussed on the interpretation of operators in different entailment contexts (O'Leary & Crain, 1994; Chierchia et al., 2001; Gualmini, 2014; Su et al., 2012; Panizza et al., 2013), such as words with logical valence or negative polarity items, we sought to investigate the *pure* phenomena of entailment relationships in a more conceptual way in children

We tested 3 to 6-year-olds, to see the evolution of the use they made of entailment to draw inferences. In parallel, we also tested adults, to have a comparison group. Our task consisted of a story involving a character that acted upon a set according to a conditional sentence. The use of a conditional sentence allowed us to study entailment contexts (UE vs. DE) with the minimum changes in the linguistic material without, since the antecedent of a conditional embeds a DE contexts and its consequent, a UE context. At the end of the story, children were asked for the subset or the superset of the target set -depending on the condition.

We found that both children and adults were sensitive to the entailment context. Children as young as three years old easily differentiated between entailment contexts and draw different inferences depending on them. This sensitivity did not differ among children of any age group, however adults' pattern was stronger than that of children.

Unexpectedly, both adults and children showed an asymmetrical pattern of computation of inferences: valid inferences in UE were much easier to draw than in DE. While for adults this asymmetry meant that their computation of valid inferences in DE was weaker than in UE, for children this meant that they did not differentiate between valid and invalid inferences in DE. The weakness of the inferences in DE was also confirmed at the individual level results.

We are currently running a control study, along with a version in Italian that seem to confirm these results also cross-linguistically. These data, along with the adult results, suggest that entailment relationships in DE contexts, in which the pattern of implication goes from the set to the subset, are less automatic than entailment relationships in the UE context, those in which the implicature goes from the set to the superset.

Interestingly, we found links of that asymmetry in studies about visual memory in both adults and infants (Halberda et al., 2006; Zosh et al., 2011). These studies suggest that while the computation of a superset (which is what has to be done in a UE valid inference) is automatic, the computation of subsets (DE valid inferences) depend on memory demands. The similarities between visual memory and linguistic entailment patterns could point at a non-linguistic origin of entailment, since this visual memory signature has been shown with preverbal infants (Zosh et al., 2011).

Our next step in this direction will be to run a second version of the study but using lexicalized concepts (<'dog', 'Dalmatian'>) instead of compositional ones (<'dog', 'dog with stains'>), to investigate the effects of the lexicaliztation in the understanding of entailment implicatures, specifically in overcoming the weakness of DE inferences. In a similar way, as the syntactic regularities of the number list help children in their learning of numbers (Chu et al., 2020), also the lexicalization of the entailment relationships could help children to achieve a symmetrical understanding of entailment relationships.

With these results in mind, we have recently started to think on non-verbal (hopefully non-linguistic) ways to investigate sensitivity to entailment, as a way to investigate it with non-verbal infants. Chapter 5 presents the results of the series of pilot experiments that we started to address this topic.

To study non-linguistic sensitivity to entailment, we first created a series of videos that featured a causal event (a hit of one or more balls that made another ball move). The group of balls allowed us to translate the quantifier (some/all of the balls), and the causal event translated a conditional sentence, which allowed us, as in the case of the study of chapter 4, to alternate between UE and DE contexts with the minimum changes on the scene.

We presented our material to adults and asked them to classify and describe them. These attempts were not especially fruitful, since eliciting responses based on a logical/quantificational dimension soon proved to be challenging. However, we found that in our participants' semi-directed descriptions their vocabulary choice was consistent with what the theory of exhaustification would predict.

These results, although weak, open the door to more ambitious projects. We believe that the fact that our participants could unconsciously capture the differences between entailment contexts in their descriptions indicates that our experimental material has the potential to elicit such differences, but that our approach was too naive in the questions we asked them.

In that sense, we intend to continue this project by exploiting the potentialities of pupil dilation paradigms (which have already been used in reasoning paradigms with reasonable success, see Cesana-Arlotti et al., 2018) to investigate the derivation of such inferences in adults first and12-months-old infants (see section 5.1) afterward.

We judge this project to be of vital importance for the field. Because the understanding of entailment relations is considered a core component of language (Su et al., 2012), and in light of our results in chapter 4 which showed that the understanding of entailment already at the age of 3, it would not seem unreasonable that infants as young as 12 months of age are capable to recognize the asymmetries that entailment relations endorse. However, on the other hand, entailment relationships are an extremely abstract relationship, which resonates in language but has its roots in logic. Finding that infants can grasp relationships with this level of abstraction would be a strong claim for the existence of propositional structures of thought before language kicks in.

6.3 Concluding remarks of a long term project

The exploration of two open problems in linguistics (the nature of words with logical valence and the role of entailment) were the door we used in this dissertation to access to the logical capacities supporting language, our first research aim. We believe that both of them provided interesting evidence pointing at a real pervasive role of logical computations in language. We judge this thesis to be a step to confirm the importance of the operation of exhaustification in language.

We are forced to be a little more speculative regarding our second aim, which was to find potential primitives of thought. However, it is worth mentioning that the operation we studied in this thesis and that we think constitute an important basis of language shares commonalities with other psychological, non-linguistic operations. We think that good candidates for primitives would be a closure operation (for which we provided evidence in chapters 2 and 3) and also the ability to estimate the informativeness of a statement and to logically conclude what it entails (which we studied its linguistic expression in chapter 4 and we tentatively investigated it non-linguistically in chapter 5).

For the first ability, a closure operation, we think that it could be not only at the basis of exhaustification but could also be in place for operations of hypothesis testing (see 1.3). Hypothesis testing needs also to restrict the space of possible hypotheses in a similar way exhaustification needs to restrict the space of alternatives. A main difference would be that hypotheses testing deals with situations or complex causes or quantification over events, while exhaustification, or at least what we have dealt with in this thesis, deals with individuals or quantification over individuals.

Actually, and rather interestingly, also the asymmetrical results about entailment we presented in chapter 4 can be read as depending on closure of the domain: participants showed a preference for identifying their domain (superset) rather than its specificities (subset). Our interest is now in taking the computations we have presented and trying to find their signature patterns into infant cognition.

Bridging the study of language and mind is crucial. The fields of language acquisition and that of developmental psychology share the target phenomena to explain: how children see and interpret the world, and why they do the way they do. And clearly when they face it they do not see a compartmentalized space, but have to deal with cognitive challenges as a whole. We think it is of vital importance for the development of the two that linguistics and psychology speak to each other and are able to come to common explanations. A deeper collaboration between the two would benefit one with empirical foundations for its theories and the other with richer theories and phenomena to explain.

We judge this project to be of vital importance. Understanding whether the building blocks that constitute language in general, and specifically exhaustificaion and entailment, are present before language emerges, would help us explain our evolution as species, contributing to the ever-lasting debate of what made us so different from the rest of the world (maybe even to explain why kangaroos do not construct rockets!). A better understanding of these building blocks would indicate not only how language is acquired since the youngest age but also, potentially, how the thought works.

For Frege (1918) and Fodor (1985a) a thought was a structured object within a structured mind. It would be hasty to state that our results conclusively advocate in favor of this assertion, however we think that they constitute a first temptative state to trace a well defined structure back to thought.

We have also our bet on this question. The world is a complex object that we have to investigate and understand from a very young age, and learning to navigate it and, specifically learning a language, is a hard work if everything has to be learned by experience. In that sense the manipulation of structured rather than unstructured entities would provide an advantage to any organism (Gualmini et al., 2003).

We are aware that we are concluding in a rather inconclusive way. The research project we started to profile in this thesis is far from being conclusive . However, we think that this endeavor is a worth attempt, whose power of explanation could shed light on where we come from, who we are and where we are going.

Appendix A

Supplementary materials

An example of the procedures of the experiments can be found at the Supplementary material folder.

Bibliography

Almoammer, A., Sullivan, J., Donlan, C., Marušič, F., O'Donnell, T., & Barner, D. (2013). Grammatical morphology as a source of early number word meanings. *Proceedings of the National Academy of Sciences*, *110*(46), 18448–18453. doi:https://doi.org/10.1073/pnas.1313652110

Ariel, M. (2015). Doubling up: Two upper bounds for scalars. *Linguistics*, *53*(3), 561–610.

Arosio, F., Foppolo, F., Pagliarini, E., Perugini, M., & Guasti, M. T. (2017). Semantic and pragmatic abilities can be spared in Italian children with SLI. *Language Learning and Development*, *13*(4), 418–429.

Barbosa, M. F. M. (2015). A Brief way on Philosophy of Language: from Plato to Port-Royal Grammar. *International Journal of Language and Literature*, *3*(1), 61–70. Retrieved from http://ijll-net.com/journals/ijll/Vol_3_No_1_June_2015/8.pdf

Barner, D. (2017). Language, procedures, and the non-perceptual origin of number word meanings. *Journal of child language*, *44*(3), 553–590. doi:https://doi.org/10.1017/S0305000917000058

Barner, D. & Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive psychology*, *60*(1), 40–62. doi:https://doi.org/10.1016/j.cogpsych.2009.06.002

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition*, *118*(1), 84–93. doi:https://doi.org/10.1016/j.cognition.2010.10.010

Barner, D., Chow, K., & Yang, S. (2009). Finding one's meaning: A test of the relation between quantifiers and integers in language development. *Cognitive Psychology*, *58*, 195–219. doi:https://doi.org/10.1016/j.cogpsych.2008.07.001

Barner, D., Hochstein, L. K., Rubenson, M. P., & Bale, A. (2018). Four-date-old children compute scalar implicatures in absence of epistemic reasoning. *Se-*

mantics in language acquisition, *24*, 325. doi:https://doi.org/10.1075/tilar.24.14bar

Barner, D., Libenson, A., Cheung, P., & Takasaki, M. (2009). Cross-linguistic relations between quantifiers and numerals in language acquisition: Evidence from Japanese. *Journal of experimental child psychology*, *103*(4), 421–440. doi:http://doi.org/10.1016/j.jecp.2008.12.001

Barner, D., Thalwitz, D., Wood, J., Yang, S.-J., & Carey, S. (2007). On the relation between the acquisition of singular–plural morpho-syntax and the conceptual distinction between one and more than one. *Developmental science*, *10*(3), 365–373. doi:https://doi.org/10.1111/j.1467-7687.2007.00591.x

Barner, D., Wood, J., Hauser, M., & Carey, S. (2008). Evidence for a non-linguistic distinction between singular and plural sets in rhesus monkeys. *Cognition*, *107*(2), 603–622. doi:http://doi.org/10.1016/j.cognition.2007.11.010

Beilin, H. & Lust, B. (1975a). Connectives: Logical, Linguistic, and Psychological Theory. In *Studies in the cognitive basis of language development* (pp. 186–216). Academic Press.

Beilin, H. & Lust, B. (1975b). A study of the development of logical and linguistic connectives: Linguistic data. In *Studies in the cognitive basis of language development* (pp. 217–284). Academic Press.

Bloom, L., Lahey, M., Hood, L., Lifter, K., & Fiess, K. (1980). Complex sentences: Acquisition of syntactic connectives and the semantic relations they encode. *Journal of child language*, *7*(2), 235–261. doi:https://doi.org/10.1017/s0305000900002610

Bod, R. (2013). *A new history of the humanities: The search for principles and patterns from antiquity to the present*. Oxford University Press.

Bonatti, L. L. & Téglás, E. (2009). *'All' and 'some' at 12 and 10 months*.

Boroditsky, L. (2006). Linguistic relativity. In *Encyclopedia of cognitive science*. John Wiley & Sons. doi:https://doi.org/10.1002/0470018860.s00567

Boroditsky, L. & Ramscar, M. (2002). The roles of body and mind in abstract thought. *Psychological science*, *13*(2), 185–189. doi:http://doi.org/10.1111/1467-9280.00434

Bott, L. & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of memory and language*, *51*(3), 437–457. doi:https://doi.org/10.1016/j.jml.2004.05.006

Braine, M. D. & O'Brien, D. P. (1998). How to Investigate Mental Logic and the Syntax of Thought. In M. Braine & D. O'Brien (Eds.), *Mental logic* (pp. 53–70). Psychology Press.

Braine, M. D. & Rumain, B. (1981). Development of comprehension of 'or': Evidence for a sequence of competencies. *Journal of experimental child psychology*, *31*(1), 46–70. doi:https://doi.org/10.1016/0022-0965(81)90003-5

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, *100*(3), 434–463. doi:https://doi.org/10.1016/j.cognition.2005.07.003

Carey, S. (2009). *The origin of concepts*. Oxford University Press. doi:https://doi.org/10.1093/acprof:oso/9780195367638.001.0001

Carey, S. & Barner, D. (2019). Ontogenetic origins of human integer representations. *Trends in cognitive sciences*, *23*(10), 823–835. doi:https://doi.org/10.1016/j.tics.2019.07.004

Carnap, R. (1968). *Logische syntax der sprache*. Springer.

Cesana-Arlotti, N., Kovács, Á. M., & Téglás, E. (2020). Infants recruit logic to learn about the social world. *Nature Communications*, *11*(1), 5999. doi:https://doi.org/10.1038/s41467-020-19734-5

Cesana-Arlotti, N., Martin, A., Téglás, E., Vorobyova, L., Cetnarski, R., & Bonatti, L. L. (2018). Precursors of logical reasoning in preverbal human infants. *Science*, *359*(6381), 1263–1266. doi:https://doi.org/10.1126/science.aao3539

Cesana-Arlotti, N., Téglás, E., & Bonatti, L. L. (2012). The probable and the possible at 12 months: Intuitive reasoning about the uncertain future. *Advances in child development and behavior*, *43*, 1–25. doi:https://doi.org/10.1016/b978-0-12-397919-3.00001-0

Chemla, E. & Spector, B. (2011). Experimental evidence for embedded scalar implicatures. *Journal of semantics*, *28*(3), 359–400.

Cheung, P., Rubenson, M., & Barner, D. (2017). To infinity and beyond: Children generalize the successor function to all possible numbers dates after learning to count. *Cognitive psychology*, *92*, 22–36. doi:https://doi.org/10.1016/j.cogpsych.2016.11.002

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. *Structures and beyond*, *3*, 39–103.

Chierchia, G. (2013). *Logic in grammar: Polarity, free choice, and intervention*. Oxford Scolarship. doi:https://doi.org/10.1093/acprof:oso/9780199697977.001.0001

Chierchia, G. (2017). Scalar implicatures and their interface with grammar. *Annual Review of Linguistics*, *3*, 245–264. doi:https://doi.org/10.1146/annurev-linguistics-011516-033846

Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., & Meroni, L. (2001). The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. In A. H.-J. Do, L. Dominguez, & A. Johansen (Eds.), *Proceedings of the 25th Boston University conference on language development* (pp. 157–168). Cascadilla Press.

Chierchia, G. & McConnell-Ginet, S. (2000). *Meaning and grammar: An introduction to semantics*. MIT press.

Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences*, *3*(1), 1–15. doi:https://doi.org/10.1017/S0140525X00001515

Chomsky, N. (1993). *Lectures on government and binding: The Pisa lectures* (7th ed.). Walter de Gruyter.

Christensen, R. H. B. (2019). ordinal—Regression Models for Ordinal Data. R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

Chu, J., Cheung, P., Schneider, R. M., Sullivan, J., & Barner, D. (2020). Counting to Infinity: Does learning the syntax of the count list predict knowledge that numbers are infinite? *Cognitive Science*, *44*(8), e12875. doi:https://doi.org/10.1111/cogs.12875

Cohen, J., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, & Computers*, *25*(2), 257–271. doi:https://doi.org/10.3758/BF03204507

Davidson, K., Eng, K., & Barner, D. (2012). Does learning to count involve a semantic induction? *Cognition*, *123*(1), 162–173. doi:https://doi.org/10.1016/j.cognition.2011.12.013

De Carvalho, A., Dautriche, I., Christophe, A., & Trueswell, J. (2019). Infants' comprehension of negative sentences while learning word meanings. In *The 44th Boston University Conference on Language Development - BUCLD*. Boston University.

Dieussaert, K., Verkerk, S., Gillard, E., & Schaeken, W. (2011). Some effort for some: Further evidence that scalar implicatures are effortful. *Quarterly Journal of Experimental Psychology*, *64*(12), 2352–2367. doi:https://doi.org/10.1080/17470218.2011.588799

Ekramnia, M. (2016). *Investigating Two Domain-General Processes in Early Infancy: Disjunctive Inference and Reorientation of Attention* (Doctoral dissertation, SISSA, Trieste, Italy). Retrieved from http://hdl.handle.net/20.500.11767/3918

Fairchild, S. & Papafragou, A. (2018). Sins of omission are more likely to be forgiven in non-native speakers. *Cognition*, *181*, 80–92.

Feigenson, L. (2011). Predicting sights from sounds: 6-month-olds' intermodal numerical abilities. *Journal of experimental child psychology*, *110*(3), 347–361. doi:https://doi.org/10.1016/j.jecp.2011.04.004

Feigenson, L. & Halberda, J. (2004). Infants chunk object arrays into sets of individuals. *Cognition*, *91*(2), 173–190. doi:https://doi.org/10.1016/j.cognition.2003.09.003

Feigenson, L. & Halberda, J. (2008). Conceptual knowledge increases infants' memory capacity. *Proceedings of the National Academy of Sciences*, *105*(29), 9926–9930. doi:https://doi.org/10.1073/pnas.0709884105

Feiman, R., Hartshorne, J. K., & Barner, D. (2019). Contrast and Entailment: Abstract logical relations constrain how 2-and 3-date-old children interpret unknown numbers. *Cognition*, *183*, 192–207. doi:https://doi.org/10.1016/j.cognition.2018.11.005

Feiman, R., Mody, S., Sanborn, S., & Carey, S. (2017). What do you mean, no? Toddlers' comprehension of logical 'no' and 'not'. *Language Learning and Development*, *13*(4), 430–450. doi:https://doi.org/10.1080/15475441.2017.1317253

Floridi, L. (2019). Semantic Conceptions of Information. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019 edition). Metaphysics Research Lab, Stanford University.

Fodor, J. A. (1985a). Fodor's guide to mental representation: The intelligent auntie's vade-mecum. *Mind*, *94*(373), 76–100. Retrieved from https://www.jstor.org/stable/2254700

Fodor, J. A. (1985b). Precis of the modularity of mind. *Behavioral and brain sciences*, *8*(1), 1–5. doi:https://doi.org/10.1017/S0140525X0001921X

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT press.

Fodor, J. A. (1989). Why there still has to be a language of thought. In P. Slezak & W. R. Albury (Eds.), *Computers, brains and minds. Australasian Studies in History and Philosophy of Science* (7th ed., pp. 23–46). Springer. doi:https://doi.org/10.1007/978-94-009-1181-9_2

Foppolo, F., Guasti, M. T., & Chierchia, G. (2012). Scalar implicatures in child language: Give children a chance. *Language learning and development*, *8*(4), 365–394. doi:https://doi.org/10.1080/15475441.2011.626386

Fox, D. (2007). Free choice and the theory of scalar implicatures. In U. Sauerland & P. Stateva (Eds.), *Presupposition and implicature in compositional semantics,Palgrave Studies in Pragmatics, Language and Cognition* (pp. 71–120). Palgrave Macmillan. doi:https://doi.org/10.1057/9780230210752_4

Frege, G. (1879). *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. L. Nebert. Retrieved from https://gallica.bnf.fr/ark:/12148/bpt6k65658c

Frege, G. (1918). The thought: A logical inquiry. *Mind*, *65*(259), 289–311. Retrieved from http://www.jstor.org/stable/2251513

Gazdar, G. (1980). Pragmatics and logical form. *Journal of Pragmatics*, *4*(1), 1–13. doi:https://doi.org/10.1016/0378-2166(80)90014-4

Gergely, G. & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in cognitive sciences*, *7*(7), 287–292. doi:https://doi.org/10.1016/S1364-6613(03)00128-1

Geurts, B. (2003). Reasoning with quantifiers. *Cognition*, *86*(3), 223–251. doi:https://doi.org/10.1016/S0010-0277(02)00180-4

Geurts, B. & van Der Slik, F. (2005). Monotonicity and processing load. *Journal of semantics*, *22*(1), 97–117. doi:https://doi.org/10.1093/jos/ffh018

Gotzner, N., Barner, D., & Crain, S. (2020). Disjunction triggers exhaustivity implicatures in 4-to 5-date-olds: Investigating the role of access to alternatives. *Journal of Semantics*, *37*(2), 219–245. doi:https://doi.org/10.1093/jos/ffz021

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Gualmini, A. (2014). *The ups and downs of child language: Experimental studies on children's knowledge of entailment relationships and polarity phenomena*. Routledge.

Gualmini, A., Meroni, L., & Crain, S. (2003). An asymmetric universal in child language. In *Proceedings of Sinn und Bedeutung* (Vol. 7, pp. 136–148).

Guasti, M. T., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and cognitive processes*, *20*(5), 667–696. doi:https://doi.org/10.1080/01690960444000250

Gweon, H. & Schulz, L. (2011). 16-month-olds rationally infer causes of failed actions. *Science*, *332*(6037), 1524–1524. doi:https://doi.org/10.1126/science.1204493

Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological science*, *17*(7), 572–576. doi:https://doi.org/10.1111/j.1467-9280.2006.01746.x

Harman, G. (2011). Quine's Semantic Relativity. *American Philosophical Quarterly*, *48*(3), 283–285. Retrieved from https://www.jstor.org/stable/23025097

Hartshorne, J. K., Snedeker, J., Liem Azar, S. Y.-M., & Kim, A. E. (2015). The neural computation of scalar implicature. *Language, cognition and neuroscience*, *30*(5), 620–634.

Havron, N., de Carvalho, A., Fiévet, A.-C., & Christophe, A. (2019). Three-to four-date-old children rapidly adapt their predictions and use them to learn novel word meanings. *Child development*, *90*(1), 82–90. doi:https://doi.org/10.1111/cdev.13113

Hochmann, J.-R., Carey, S., & Mehler, J. (2018). Infants learn a rule predicated on the relation same but fail to simultaneously learn a rule predicated on the relation different. *Cognition*, *177*, 49–57. doi:https://doi.org/10.1016/j.cognition.2018.04.005

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. doi:https://doi.org/10.1016/j.cogpsych.2016.01.005

Hochstein, L., Bale, A., & Barner, D. (2018). Scalar implicature in absence of epistemic reasoning? The case of autism spectrum disorder. *Language Learning and Development*, *14*(3), 224–240.

Hodges, W. (2001). *Logic-An Introduction to Elementary Logic* (2nd ed.). Penguin books London.

Horn, L. (1989). *A natural history of negation*. University of Chicago Press.

Horn, L. R. (2006). The border wars: A neo-Gricean perspective. In K. Heusinger & K. Turner (Eds.), *Where semantics meets pragmatics* (Vol. 16, pp. 21–48). Oxford: Elsevier. Retrieved from https://ling.yale.edu/sites/default/files/files/horn/horn05_borderwars.pdf

Horn, L. R. (1972). *On the semantic properties of logical operators in English* (Doctoral dissertation, University of California, Los Angeles). Retrieved from https://linguistics.ucla.edu/images/stories/Horn.1972.pdf

Huang, Y. T. & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive psychology*, *58*(3), 376–415. doi:https://doi.org/10.1016/j.cogpsych.2008.09.001

Huang, Y. T., Spelke, E., & Snedeker, J. (2013). What exactly do numbers mean? *Language Learning and Development*, *9*(2), 105–129. doi:https://doi.org/10.1080/15475441.2012.658731

Hurewitz, F., Papafragou, A., Gleitman, L., & Gelman, R. (2006). Asymmetries in the acquisition of numbers and quantifiers. *Language learning and development*, *2*(2), 77–96. Retrieved from http://papafragou.psych.udel.edu/papers/asymmetries.pdf

Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences*, *106*(25), 10382–10385. doi:https://doi.org/10.1073/pnas.0812142106

Janssen, T. M. V. (2020). Montague Semantics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020 edition). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/entries/montague-semantics/

Jasbi, M. & Frank, M. C. (2017). The Semantics and Pragmatics of Logical Connectives: Adults' and Children's Interpretations of And and Or in a Guessing Game. In *39th Annual Conference Cognitive Science Society*. Retrieved from https://cogsci.mindmodeling.org/2017/papers/0117/paper0117.pdf

Jorgensen, J. C. & Falmagne, R. J. (1992). Aspects of the meaning of if-then for older preschoolers: Hypotheticality, entailment, and suppositional processes. *Cognitive Development*, *7*(2), 189–212. doi:https://cogsci.mindmodeling.org/2017/papers/0117/paper0117.pdf

Kabdebon, C., Pena, M., Buiatti, M., & Dehaene-Lambertz, G. (2015). Electrophysiological evidence of statistical learning of long-distance dependencies in 8-month-old preterm and full-term infants. *Brain and language*, *148*, 25–36. doi:https://doi.org/10.1016/j.bandl.2015.03.005

Katsos, N. & Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, *120*(1), 67–81.

Kayhan, E., Gredebäck, G., & Lindskog, M. (2018). Infants distinguish between two events based on their relative likelihood. *Child Development*, *89*(6), e507–e519. doi:https://doi.org/10.1111/cdev.12970

Knowlton, T., Pietroski, P., Halberda, J., & Lidz, J. (2019). The Mental Representation of Universal Quantifiers. Retrieved from https://ling.auf.net/lingbuzz/004486

Ladusaw, W. (1979). *Negative polarity items as inherent scope relations* (Doctoral dissertation, University of Texas at Austin).

Lakoff, G. (1969). On generative semantics. *Semantics- An Interdisciplinary reader in Philosophy, Linguistics, Anthropology and Psychology*, 232–296.

Le Corre, M. & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, *105*(2), 395–438. doi:https://doi.org/10.1016/j.cognition.2006.10.005

Li, P., Le Corre, M., Shui, R., Jia, G., & Carey, S. (2003). Effects of plural syntax on number word learning: A cross-linguistic study. In *28th Boston University Conference on Language Development, Boston, MA*.

Libertus, M. E., Feigenson, L., & Halberda, J. (2018). Infants extract frequency distributions from variable approximate numerical information. *Infancy*, *23*(1), 29–44. doi:https://doi.org/10.1111/infa.12198

Linebarger, M. C. (1980). *The grammar of negative polarity* (Doctoral dissertation, Massachusetts Institute of Technology).

Lust, B. & Mervis, C. A. (1980). Development of coordination in the natural speech of young children. *Journal of Child Language*, *7*(2), 279–304. doi:https://doi.org/10.1017/S0305000900002634

Marchetto, E. & Bonatti, L. L. (2013). Words and possible words in early language acquisition. *Cognitive psychology*, *67*(3), 130–150. doi:https://doi.org/10.1016/j.cogpsych.2013.08.001

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80. doi:https://doi.org/10.1126/science.283.5398.77

Marty, P., Chemla, E., & Spector, B. (2013). Interpreting numerals and scalar items under memory load. *Lingua*, *133*, 152–163.

Marušič, F., Žaucer, R., Saksida, A., Sullivan, J., Skordos, D., Wang, Y., & Barner, D. (2020). Do children derive exact meanings pragmatically? Evidence from a dual morphology language. *Cognition*, *207*, 104527–104527. doi:https://doi.org/10.1016/j.cognition.2020.104527

Marušič, F., Plesničar, V., Razboršek, T., Sullivan, J., Barner, D., et al. (2016). Does grammatical structure accelerate number word learning? Evidence from learners of dual and non-dual dialects of Slovenian. *PloS one*, *11*(8), e0159208. doi:https://doi.org/10.1371/journal.pone.0159208

McGonigle, B. & Chalmers, M. (1996). The ontology of order. In L. Smith (Ed.), *Critical readings on Piaget* (pp. 279–311). Routledge.

Menzies, P. & Beebee, H. (2020). Counterfactual Theories of Causation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 edition). Metaphysics Research Lab, Stanford University.

Mody, S. & Carey, S. (2016). The emergence of reasoning by the disjunctive syllogism in early childhood. *Cognition*, *154*, 40–48. doi:https://doi.org/10.1016/j.cognition.2016.05.012

Montague, R. (1970). Universal grammar. *Theoria*, *36*(3), 373–398. doi:https://doi.org/10.1111/j.1755-2567.1970.tb00434.x

Morris, B. J. (2008). Logically speaking: Evidence for item-based acquisition of the connectives AND & OR. *Journal of Cognition and Development*, *9*(1), 67–88. doi:https://doi.org/10.1080/15248370701836600

Nadel, L. & Piattelli-Palmarini, M. (2003). What is cognitive science. In *Encyclopedia of cognitive science, London: Macmillan*. Macmillan. Retrieved from http://www.biolinguistics.uqam.ca/Nadel&Piattelli-Palmarini_2003.pdf

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188. doi:https://doi.org/10.1016/s0010-0277(00)00114-1

O'Leary, C. & Crain, S. (1994). Negative polarity items (a positive result) positive polarity items (a negative result). In *Boston University Conference on Language Development*.

Pagliarini, E., Bill, C., Romoli, J., Tieu, L., & Crain, S. (2018). On children's variable success with scalar inferences: Insights from disjunction in the scope of a universal quantifier. *Cognition*, *178*, 178–192. doi:https://doi.org/10.1016/j.cognition.2018.04.020

Pagliarini, E., Crain, S., & Guasti, M. T. (2018). The Compositionality of Logical Connectives in Child Italian. *Journal of psycholinguistic research*, *47*(6), 1243–1277. doi:https://doi.org/10.1007/s10936-018-9596-1

Panizza, D., Chierchia, G., & Clifton Jr, C. (2009). On the role of entailment patterns and scalar implicatures in the processing of numerals. *Journal of memory and language*, *61*(4), 503–518. doi:https://doi.org/10.1016/j.jml.2009.07.005

Panizza, D., Huang, Y. T., Chierchia, G., & Snedeker, J. (2009). Relevance of polarity for the online interpretation of scalar terms. In *Semantics and Linguistic Theory (SALT)* (Vol. 19, pp. 360–378). doi:https://doi.org/10.3765/salt.v19i0.2530

Panizza, D., Notley, A., Thornton, R., & Crain, S. (2013). When children are as logical as adults: the interpretation of numerals in child language. In S. Baiz, N. Goldman, & R. Hawkes (Eds.), *Boston University Conference on Language Development (BUCLD 37)* (pp. 306–318). Cascadilla Press.

Papafragou, A. & Musolino, J. (2003). Scalar implicatures: experiments at the semantics–pragmatics interface. *Cognition*, *86*(3), 253–282. doi:https://doi.org/10.1016/s0010-0277(02)00179-8

Paris, S. G. (1973). Comprehension of language connectives and propositional logical relationships. *Journal of experimental child psychology*, *16*(2), 278–291. doi:https://doi.org/10.1016/0022-0965(73)90167-7

Partee, B. H. (1984). Nominal and temporal anaphora. *Linguistics and philosophy*, *7*(3), 243–286. Retrieved from https://www.jstor.org/stable/25001168

Partee, B. H. (2004). Reflections of a formal semanticist. Blackwell Publishers.

Piaget, J. & Inhelder, B. (1951). *The origin of the idea of chance in children*. Psychology Press, Taylor & Francis Group. doi:https://doi.org/10.4324/9781315766959

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2), 199–217. doi:https://doi.org/10.1016/j.cognition.2011.11.005

Politzer, G. (2016). The class inclusion question: a case study in applying pragmatics to the experimental study of cognition. *SpringerPlus*, *5*(1), 1–20. doi:https://doi.org/10.1186/s40064-016-2467-z

Pouscoulous, N., Noveck, I. A., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language acquisition*, *14*(4), 347–375. doi:https://doi.org/10.1080/10489220701600457

Pulvermüller, F. (2013). Semantic embodiment, disembodiment or misembodiment? In search of meaning in modules and neuron circuits. *Brain and language*, *127*(1), 86–103. doi:https://doi.org/10.1016/j.bandl.2013.05.015

Redshaw, J. & Suddendorf, T. (2016). Children's and apes' preparatory responses to two mutually exclusive possibilities. *Current Biology*, *26*(13), 1758–1762. doi:https://doi.org/10.1016/j.cub.2016.04.062

Rives, B. (2020). Fodor, Jerry. In *Internet Encyclopedia of Philosophy* (2020 edition). Fieser, James and Dowden, Bradley. Retrieved from https://iep.utm.edu/fodor/

Rooth, M. (1992). A theory of focus interpretation. *Natural language semantics*, *1*(1), 75–116. doi:https://doi.org/10.1007/BF02342617

Rosenberg, R. D. & Feigenson, L. (2013). Infants hierarchically organize memory representations. *Developmental science*, *16*(4), 610–621. doi:https://doi.org/10.1111/desc.12055

Russell, B. (1905). On denoting. *Mind*, *14*(56), 479–493. Retrieved from https://www.uvm.edu/~lderosse/courses/lang/Russell(1905).pdf

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and philosophy*, *27*(3), 367–391. doi:https://doi.org/10.1023/B:LING.0000023378.71748.db

Schneider, R. M., Pankonin, A. H., Schachner, A., & Barner, D. (2021). Starting small: Exploring the origins of successor function knowledge. *Developmental Science*, e13091. doi:https://doi.org/10.1111/desc.13091

Schneider, R. M., Sullivan, J., Guo, K., & Barner, D. (2021). What counts? Sources of knowledge in children's acquisition of the successor function. *Child Development*. doi:https://doi.org/10.1111/cdev.13524

Scholnick, E. K. & Wing, C. S. (1995). Logic in conversation: Comparative studies of deduction in children and adults. *Cognitive Development*, *10*(3), 319–345. doi:https://doi.org/10.1016/0885-2014(95)90001-2

Schulz, P. & Roeper, T. (2011). Acquisition of exhaustivity in wh-questions: A semantic dimension of SLI? *Lingua*, *121*(3), 383–407. doi:https://doi.org/10.1016/j.lingua.2010.10.005

Sedley, D. (2018). Plato's Cratylus. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2018). Metaphysics Research Lab, Stanford University.

Seidenberg, M. S. & Petitto, L. A. (1979). Signing behavior in apes: A critical review. *Cognition*, *7*(2), 177–215. doi:https://doi.org/10.1016/0010-0277(79)90019-2

Singh, R., Wexler, K., Astle, A., Kamawar, D., & Fox, D. (2013). Children interpret disjunction as conjunction: consequences for the theory of scalar implicatures. *Carleton University, ms.*

Skordos, D. & Papafragou, A. (2016). Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition*, *153*, 6–18. doi:https://doi.org/10.1016/j.cognition.2016.04.006

Spector, B. (2007). Scalar implicatures: Exhaustivity and gricean reasoning. In *Questions in dynamic semantics* (pp. 225–249). Brill.

Stahl, A. E. & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94. doi:https://doi.org/10.1126/science.aaa3799

Stahl, A. E. & Feigenson, L. (2017). Expectancy violations promote learning in young children. *Cognition*, *163*, 1–14. doi:https://doi.org/10.1016/j.cognition.2017.02.008

Stiller, A., Goodman, N., & Frank, M. (2011). Ad-hoc scalar implicature in adults and children. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, *33*). Retrieved from http://web.stanford.edu/~ngoodman/papers/SGF-cogsci2011.pdf

Su, Y. E., Zhou, P., & Crain, S. (2012). Downward entailment in child Mandarin. *Journal of child language*, *39*(5), 957–990. doi:https://doi.org/10.1017/S0305000911000389

Téglás, E. & Bonatti, L. L. (2016). Infants anticipate probabilistic but not deterministic outcomes. *Cognition*, *157*, 227–236. doi:https://doi.org/10.1016/j.cognition.2016.09.003

Téglás, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences*, *104*(48), 19156–19159. doi:https://doi.org/10.1073/pnas.0700271104

Tieu, L., Romoli, J., Zhou, P., & Crain, S. (2016). Children's knowledge of free choice inferences and scalar implicatures. *Journal of Semantics*, *33*(2), 269–298. doi:https://doi.org/10.1093/jos/ffv001

Tomlinson Jr, J. M., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of memory and language*, *69*(1), 18–35. doi:https://doi.org/10.1016/j.jml.2013.02.003

Van Orman, Q. W. (1960). Word and Object. *Massachusetts Institute of Technology*.

Wagner, K., Chu, J., & Barner, D. (2019). Do children's number words begin noisy? *Developmental science*, *22*(1), e12752. doi:https://doi.org/10.1111/desc.12752

Wagner, K., Kimura, K., Cheung, P., & Barner, D. (2015). Why is number word learning hard? Evidence from bilingual learners. *Cognitive psychology*, *83*, 1–21. doi:https://doi.org/10.1016/j.cogpsych.2015.08.006

Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155–193. doi:https://doi.org/10.1016/0010-0277(90)90003-3

Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive psychology*, *24*(2), 220–251. doi:https://doi.org/10.1016/0010-0285(92)90008-P

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, *85*(3), 223–250. doi:ttps://doi.org/10.1016/S0010-0277(02)00109-9

Xu, F. (2019). Towards a rational constructivist theory of cognitive development. *Psychological review*, *126*(6), 841. doi:https://doi.org/10.1037/rev0000153

Zalta, E. N. (2020). Gottlob Frege. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/fall2020/entries/frege/

Zehr, J. & Schwarz, F. (2018). PennController for Internet Based Experiments (IBEX). doi:https://doi.org/10.17605/OSF.IO/MD832

Zosh, J. M., Halberda, J., & Feigenson, L. (2011). Memory for multiple visual ensembles in infancy. *Journal of Experimental Psychology: General*, *140*(2), 141. doi:https://doi.org/10.1037/a0022925