Arnold Kochari • **Perceiving and communicating magnitudes**

Perceiving and communicating magnitudes:

Behavioural and electophysiological studies

Arnold Kochari

Perceiving and communicating magnitudes: Behavioral and electrophysiological studies.

Arnold Kochari

Perceiving and communicating magnitudes: Behavioral and electrophysiological studies.

ILLC Dissertation Series DS-2020-10

Institute for Logic, Language and Computation

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

This research was supported by the Netherlands Organization for Scientific Research (NWO) grant *Language in Interaction*, grant no. 024.001.006.

Cover design by Linnea Colliander Celik, inspired by Verena Loewensberg's art.
Printed and bound by Ipskamp Printing.

ISBN: 978-94-028-2126-0

Perceiving and communicating magnitudes: Behavioral and electrophysiological studies.

Academisch Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen
op donderdag 17 september 2020, te 10.00 uur

door

Arnold Kochari

geboren te Qabala

Promotiecommisie

Promotor:	Prof. dr. ing. R.A.M. van Rooij	Universiteit van Amsterdam
Promotor:	Prof. dr. H.J. Schriefers	Radboud Universiteit Nijmegen
Co-promotor:	Dr. J.K. Szymanik	Universiteit van Amsterdam
Overige leden:	Prof. dr. F.J.M.M. Veltman	Universiteit van Amsterdam
	Prof. dr. M. van Lambalgen	Universiteit van Amsterdam
	Prof. dr. S.J.L. Smets	Universiteit van Amsterdam
	Prof. dr. J.M. McQueen	Radboud Universiteit Nijmegen
	Dr. W.H. Zuidema	Universiteit van Amsterdam
	Dr. S. Solt	Leibniz-Centre General Linguistics

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Bavay saal nanay s'iyen śampezu.

Contents

Acknowledgments

The past 4 years of work on this thesis (and other projects) have been filled with ups and downs. Alternatingly, I have either thought of myself as the luckiest person in the world to have this job or seriously considered quitting. Nonetheless, I had a lot of fun, and I have learnt a lot. I am fortunate to have been surrounded by amazing teachers, mentors, and friends over the years.

First, I am grateful to my thesis supervisors for believing in me enough to hire me. During the project, they gave me a great deal of freedom to pursue the research questions which I myself found interesting. Robert was always available on short notice to answer any of my questions, both those which were simple and those which were foundational. I also greatly enjoyed collaborating with him on several side projects. Herbert took the time to not only give me general supervision but also to brainstorm project ideas and to discuss the details pertaining to my experiments and analyses. I would always leave our long meetings with much clearer ideas and a greater excitement for my projects. Herbert also took the time to give me detailed comments on my writing and to guide me through important conference submissions and presentations. Last but not least, Jakub helped me navigate the different scientific fields which I was venturing into and his never-fading passion for research was a huge inspiration to me throughout these years (and this is not an exaggeration). On many occasions, Jakub also provided me with invaluable advice on matters related to my professional development. Each of my three advisors contributed a great deal to shaping me as a researcher, for which I am immensely grateful.

I would also like to thank my doctorate committee members. I am honored that you accepted the invitation to be in the committee, and I am grateful for the time that you spent reading and evaluating my (occasionally digressive) thesis.

Throughout my PhD, the ILLC office has taken excellent care of all the administrative business that went into my work. I am especially grateful to Jenny Batson for quickly solving any issues with which I came to her.

There are two more people whom I would like to thank for their academic

guidance during my PhD. Firstly, I am grateful to Monique Flecken, who supervised a project that I worked on as my master's thesis. I continued working on that project with Monique in parallel to the PhD during the first couple of years. Her trust in me throughout that time gave me the self-confidence I needed to form and voice my own opinions on matters of science. Monique was also the first to show me how science works backstage — my first submitted article, first peer-review process, and first discussions about effect sizes of interest and replicability issues happened under her supervision.

Secondly, I must thank Ashley Lewis, with whom I worked on the large MEG project presented in chapter 5 of this thesis. Over a period of a year and a half, Ashley dedicated a lot of his time to teaching me about MEG data collection and analysis; discussing details of the experimental procedures and analyses with me; patiently answering all of my questions; and helping me improve my Matlab scripts, presentations and the corresponding paper. Ashley, thank you for being such a dedicated mentor and for being such a good friend.

I would also like to express my gratitude to those people whom I had the fortune of meeting in the years of education leading up to this PhD project. These people played a crucial role in shaping me as a scientist and as a person.

My interest in scientific research was born when I studied for my bachelor's degree at the Faculty of Humanities at Charles University in Prague. Zdeněk Pinc who gave introductory philosophy classes is the person who first awoke my scientific curiosity. The group of teachers I had while studying there gave me a very strong foundation in research methods which I — to this day — employ in my scientific and everyday reasoning.

My first encounter with psycholinguistics took place at Amsterdam University College in a class with Tom Lentz. I enjoyed our class assigned research project so much that I immediately decided that it was the field in which I wanted to pursue a career. Thereafter Filip Smolík in Prague agreed to supervise my bachelor's thesis and gave me a valuable opportunity to design and collect data for my project in a real psycholinguistics lab. Filip is also the person who first showed me R code - it was love at first sight.

Another person who probably does not even remember me but played a huge role in my life is Martin Everaert. I met Martin briefly when taking a short summer course, and he told me I should consider applying for the research master's programme in linguistics at Utrecht University. Until then, I did not think I could ever be admitted to a programme like that – everyone there seemed too smart, too rich, too fancy for me to even be in the same room. To my surprise, not only was I admitted but I was also awarded the Utrecht Excellence Scholarship which covered my tuition fees and living expenses, something which I could have never afforded on my own. For this generous gift I will always be grateful to the Utrecht University Alumni Fund. Words cannot describe how happy I was with the teachers and researchers I met during my master's studies at UiL OTS. Nor can I single anyone out, since every single person which I had the fortune

to meet taught me a lot. It was a fantastic experience. As part of my master's studies, I also had the fortune to spend several months working with Olaf Hauk at MRC CBU in Cambridge. Even though it was a short visit, Olaf made a great contribution to the set of skills and knowledge that I now have.

Having thus thanked my academic mentors, I would also like to briefly thank my colleagues and friends. Thanks to Adam, with whom I have had many long discussions ever since the beginning of our friendship in Prague. Thanks to Florian for his companionship across various libraries over the years — I have the warmest memories of those days. Thanks to Shuangshuang, Rachel, and Suzanne (and later also Nick), my study-buddies in Utrecht with whom I quickly became good friends. Thanks to Markus, Nadine and Limor in Nijmegen, with whom I shared many pleasant discussions, lunches and dinners. Markus was also nice enough to read and provide thoughtful feedback on some texts in this thesis. In Amsterdam, I must thank my colleagues-turned-friends Bastiaan, Dean, Dieuwke, Frederik, Iris, Jasmijn, Levin, Mostafa, Nadine, Ronald, Samira, and Thomas. I am especially grateful to those friends who were able to host me while I was collecting data in Nijmegen and, after having moved to Stockholm, while I was visiting the Netherlands. For the last couple of months of work, I must thank my thesis-writing buddy Victor who helped me stay focused and with whom I shared many hours of commiserating.

Finally, I must extend my greatest gratitude towards my family. Thanks to my parents and my sister for their unwavering faith in me and support of me as I pursued my dreams, despite how far away from home and family it has brought me. Больше всех за успех я обязан своей семье. Я благодарен своим родителям и сестре за то, что они всегда в меня верили, за их нескончаемую поддержку, и за то, что они позволили мне следовать своей мечте хоть это и означало, что я уезжал далеко от них. I am also grateful to my now wife Linnea, for providing unlimited emotional support throughout my PhD, with all its ups and downs. Thanks for helping me in every aspect, for forgiving the numerous stressful evenings and weekends which I have spent working and not with you, for proofreading the most important bits of this thesis, and for creating a very fitting cover design. I could never dare to attempt something like this without the support of my family. Я бы никогда не осмелился подумать, что я смогу заниматься исследованиями и написать докторскую диссертацию без поддержки моей семьи.

Stockholm
July, 2020.

Arnold Kochari

Chapter 1

Introduction

It is a sunny day in Bos en Lommer, and I am waiting for my friend at our favorite pizza place. As soon as as we spot each other, we wave, and before my friend has locked their bike, they exclaim:

(1) I just saw a *large rabbit* on my way here!

or:

(2) I just saw *many rabbits* on my way here!

Among other things, what has happened here is that my friend has observed and estimated a size of a rabbit or a quantity of rabbits and is now communicating this information about the observed size or quantity to me using natural language. The ability to make estimates of and compare magnitudes along various perceptual dimensions such as size, quantity, length, loudness, weight, duration, etc. is one of the fundamental cognitive capacities that helps humans navigate the world. As social creatures equipped with a complex language processing system, we are able to communicate information about our mental states to each other, which includes communicating magnitudes. In the first sentence above, information about size was communicated using the word 'large', a *scalar adjective*. Other examples of scalar adjectives in English are 'big', 'small', 'long', 'short', 'loud', 'quiet', 'high', 'low', etc. In the second sentence, information about quantity was communicated using 'many', a *vague quantifier*. Other vague quantifiers in English are 'few', 'much', and 'little'. Scalar adjectives and vague quantifiers are some of the most frequent words in everyday language,[1] and they play a major role in successful communication. The present thesis is concerned with the cognitive mechanisms underlying the use of these types of words.

[1]Eight of the words given as examples here are among the 500 most frequent words in British English, and all are among the 2500 most frequent words in British English (based on frequencies in SUBTLEX-UK corpus; van Heuven, Mandera, Keuleers, & Brysbaert, 2014)

How do we understand the size or the quantity conveyed by the words 'large' and 'many' in the sentences above? I will explain this with the example of the scalar adjective, but this discussion is equally valid for vague quantifiers.

One distinctive property of scalar adjectives in comparison to other words in human languages is that they can in fact be used to describe a vast range of magnitudes. For example, a 'large bee', a 'large lion', or a 'large whale' are all acceptable uses of 'large', yet they describe objects that are widely different in size. Understanding the noun with which an adjective like 'large' is combined is a necessary prerequisite for the interpretation of the magnitude described by such a scalar adjective. In addition, the interpretation of what my friend said will also depend on the common size of rabbits in Amsterdam, my previous experiences with rabbits of different sizes, what kind of rabbits I think my friend has seen, what my friend wanted to achieve by saying this phrase to me, and other factors. Given that all of these factors are important, it is obvious that the meaning of scalar adjectives — and of vague quantifiers — to a large extent depends on the context in which they are used. This is the first property that is of interest in this thesis.

However, knowing these aspects of the context will still not be enough to determine the exact size of the rabbit that my friend has just seen. This is the case because scalar adjectives in fact always refer to approximate rather than exact magnitudes. When we see two rabbits of approximately the same size, we cannot say that one of them is 'large' whereas the other is 'not large'. Perhaps we can even see that one of the rabbits is slightly larger than the other, but because the difference is small, if we say that one of them is 'large' we have to say that the other one is 'large' too. This demonstrates that 'large' does not allow for drawing sharp distinctions between magnitudes. The fact that scalar adjectives — as well as vague quantifiers — refer to approximate magnitudes that do not allow for drawing sharp distinctions is second property of interest in this thesis.

This thesis consists of a series of studies investigating the cognitive and neuronal processes that take part in the comprehension and production of scalar adjectives and quantifiers. These investigations are based upon, contribute to, and hopefully bridge research in semantics, numerical cognition, psycholinguistics, and the cognitive neuroscience of language. Each chapter has been written as a self-contained manuscript and can thus be read independently. Nonetheless, there is a unifying theme: two specific properties of scalar adjectives and vague quantifiers, namely context-dependence and the imprecise nature of the magnitudes that they describe. In the next sections, I will introduce quantifiers, scalar adjectives, and methodological considerations in more detail and give some background information regarding each of the chapters of the thesis.

1.1 Properties of the meaning of scalar adjectives and vague quantifiers

Let us consider the properties of the meaning of scalar adjectives and vague quantifiers in more detail. Here, I give a general overview of the properties that have been the subject of a long tradition of research in formal semantics (see e.g., Kennedy, 2007; Lassiter, 2015; Solt, 2011, 2015a, 2015b; van Rooij, 2011b). Research on semantics is not (normally) concerned with processing questions, but this thesis takes the insights gained from semantics into account because they can provide useful information on (restrictions of) processing.

Adjectives such as 'large', 'small', 'long', 'short', 'loud', 'quiet', etc., that I here call *scalar adjectives* are gradable, context-sensitive, and vague.[2] The vague quantifiers 'many', 'few', 'much', and 'little' share these properties. The discussion in this section is thus applicable to both scalar adjectives and vague quantifiers. Chapter 2 of the present thesis discusses other classes of quantifiers along with the ones mentioned here, but to keep things simple, here I only discuss the properties of vague quantifiers.

Gradable adjectives are contrasted with *non-gradable adjectives*. Whereas gradable adjectives describe continuous properties for which there are many ordered degrees (e.g., an object can be of many different sizes), non-gradable adjectives refer to properties that are by default not continuous or ordered. For example, 'pregnant', 'even', 'dead', 'rectangular', 'boiling', and 'wooden' are non-gradable. An easy way to tell whether an adjective is gradable or non-gradable is by considering whether this adjective can be combined with 'very' (e.g., 'very large' vs. ?'very dead') and whether it forms comparative and superlative forms (e.g., 'larger', 'largest' vs. ?'more dead', ?'most dead'). Another important difference is that the meaning of gradable adjectives largely depends on context, whereas the meaning of non-gradable adjectives is by default identical across different contexts.[3] Within the class of gradable adjectives, *relative* adjectives (the class of adjectives to which all the adjectives I gave as examples so far belong) are often distinguished from *absolute* adjectives such as 'full', 'empty', 'wet', 'dry', 'clean', etc. According to some proposals, unlike the relative adjectives, absolute adjectives are not context-sensitive to the same extent (i.e., they do not require the computation of a threshold (see below) given the context in which the phrase

[2]Note that in the chapters that make up this thesis I have used the term *scalar* (rather than *gradable* or *vague*) to refer to these adjectives, as this particular term is used in research on these adjectives in psycholinguistics.

[3]Here and above I say 'by default' because it is still possible to use the imagination and stretch the meaning of each of the adjectives to make the property that they describe continuous. For example, one could arguably say that someone whose baby is due in two weeks is 'more pregnant' than someone whose is due in six months. This is more of a general property of language, since in this way it is also possible to stretch the meaning of nouns (e.g., a person who is 20 years old may be referred to as a 'child' when compared to a person who is 80 years old).

is used; Kennedy, 2007; Kennedy & McNally, 2005, but see Lassiter & Goodman, 2013 and Qing & Franke, 2014 for different suggestions). This thesis is mainly concerned with relative gradable adjectives.

Scalar adjectives and vague quantifiers are also *context-sensitive*. The meaning of scalar adjectives largely depends on the noun that they are combined with. It also depends on speaker and listener experiences and expectations and possibly a number of other factors. In formal semantics, it is suggested that the noun provides a *comparison class*, a set of possible magnitudes, and that understanding the meaning of the scalar adjective requires the computation of a *threshold* or *standard of comparison*, which the object needs to meet in order to be described with the given scalar adjective (e.g., Graff, 2000; Kennedy, 2007; Kennedy & McNally, 2005; E. Klein, 1980; van Rooij, 2011a, among others). Given the noun, the listener needs to determine which comparison class is appropriate (it can be more or less ambiguous depending on how restrictive the noun phrase that the adjective modifies is) and which threshold should be applied. In the example sentence with a 'large rabbit' above, the comparison class consists of sizes of other rabbits (possibly only rabbits in Amsterdam), and the threshold is some size beyond which they would be considered 'large'. The expectations of the listener and the speaker, as well as other factors, are taken into account in determining the comparison class and the threshold. Recent modelling work proposes that the threshold is determined probabilistically from the distribution of the degrees in the comparison class such that it maximizes communicative efficiency — a trade-off between what is likely to be true and the informativeness value of the utterance (see proposals by Lassiter & Goodman, 2013; Qing & Franke, 2014; for a recent review of other proposals for threshold computation and experimental research investigating this question, see Solt, 2019).

Another property of the meaning of scalar adjectives and vague quantifiers that is generally agreed upon is a lack of sharp boundaries in their meaning / applicability — *vagueness* (see e.g., Alxatib & Sauerland, 2019; Graff, 2000; Solt, 2015b; van Rooij, 2011b for extensive overviews of vagueness). Specifically, even in a particular context, there will still be borderline cases — objects for which it is unclear whether the relevant scalar adjective applies or not, e.g. objects for which it is unclear whether they are 'large' or 'not large'. For example, an adult person who has a height of 190 cm will clearly be considered 'tall' in a Western European context, whereas an adult person who has a height of 150 cm will clearly be considered 'not tall'. At the same time, a person who has a height of 170 cm is neither clearly 'tall' nor clearly 'not tall'; this person is a borderline case of applicability of the scalar adjective 'tall'. Another phenomenon that arises due to vagueness is the so-called sorites paradox. Imagine a situation in which we see a person who has a height of 190 cm, and we decide that in this context this person can be considered 'tall' (through determination of the comparison class and computation of the threshold). Now we see a person who has a height of 189 cm next to the first person. Because very small differences

in height do not seem to matter for the applicability of 'tall', we would agree that this person should also be described as 'tall'. By this reasoning, if we keep considering additional people, each of them only 1 cm shorter than the previous one, at some point we would have to say that a person who has a height of 140 cm is also 'tall', a conclusion that is clearly counter-intuitive. This contradiction, that the last person is concluded to be 'tall' when in fact this person cannot be considered 'tall', constitutes the paradox. Borderline cases and the sorites paradox illustrate that vague expressions do not give rise to clear and accessible cut-off points as their threshold; the computed threshold does not clearly divide objects into those that can be accurately described with a scalar adjective and those that cannot. Many explanations for why the sorites paradox arises have been proposed (see e.g., D. Hyde & Raffman, 2018; van Rooij, 2011b). One type of explanation of the paradox (and the vagueness of scalar adjectives and vague quantifiers) is related to the limits of our perception of magnitudes — these words do not refer to magnitudes with strict cut-off points because we are not normally able to easily distinguish small differences within the limits of our perceptual system (Wright, 1975). This explanation will be of special interest in this thesis; it will be discussed in more detail below.

1.2 Quantifiers and number symbols as symbolic references to magnitude information

So far, I have discussed the vague quantifiers 'many', 'few', 'much', and 'little', which are most similar to scalar adjectives in terms of the properties described above (Solt, 2011, 2015a). Several other classes of quantifiers can be distinguished (Paperno & Keenan, 2017; Peters & Westerståhl, 2006; Szymanik, 2016b) some of which also share these properties, while others do not. For example, quantifiers like 'some', 'several', 'most', and 'a few' are arguably also context-sensitive and vague, but perhaps not to the same extent. Other quantifiers, such as 'all', 'every', and 'each' are not generally considered to be vague or context-sensitive. One approach to understanding how our brain stores and manipulates quantifiers (as well as scalar adjectives, discussed in detail in the next subsection) is to investigate whether and how their meaning interacts with the cognitive system for estimating and comparing the perceptual magnitudes. Before discussing this further, I will briefly introduce what we know about the human perception of numerical magnitude (i.e., quantity, the number of distinct elements)

Humans possess the ability to estimate quantities and operate with these estimated quantities (such as comparison to another quantity) as well as the ability to count and operate with number symbols referring to exact quantities. The ability to estimate and compare perceptually presented quantities appears to be present in all humans and to be innate. Prelinguistic infants and other animals have this ability too (e.g., Brannon & Terrace, 1998; Cantlon & Brannon, 2007;

Izard, Dehaene-Lambertz, & Dehaene, 2008; Xu & Spelke, 2000). Estimations of quantities that we make are noisy (e.g., we may estimate that we are seeing 7, 8, 9, 10, 11 objects when the actual number of objects is exactly 9; we are more likely to give an estimate closer to the real quantity than an estimate further way from the real quantity) and the amount of noise increases the larger the quantity to be estimated. In case of comparison, our performance in terms of accuracy and reaction time depends on the ratio between any two quantities rather than the absolute quantity of items in these sets (Feigenson, Dehaene, & Spelke, 2004; Halberda & Feigenson, 2008). Specifically, the larger the ratio between quantities (i.e., the farther apart the quantities are), the better our performance will be. This pattern is present in both humans and other animals.

Let us now take a look at number symbols. My friend could have told me:

(3) I just saw *five rabbits* on my way here!

We know that exact number symbols such 'one', 'two', 'five', 'six', etc. and operations with them need to be explicitly taught, since not all languages have an extensive numerical system. Some languages have number words only up to a certain quantity (they use what appear to be imprecise quantifiers to refer to other quantities; Bowern & Zentz, 2012; Pica, Lemer, Izard, & Dehaene, 2004). Thus, exact number symbols constitute a cultural invention.

A lot of research in the field of numerical cognition has been devoted to understanding how humans learn, store, and manipulate exact number symbols, as opposed to how they do so for approximate quantities, from perceptual input. It has been suggested that number symbol representations and processing mechanisms overlap with representations and cognitive mechanisms that initially evolved in humans for nonsymbolic quantities (Dehaene & Cohen, 2007; Nieder, 2016). On some of the suggestions, children learn number symbols by associating them with approximate quantities; it has furthermore been suggested that approximate quantity representations get recruited for the comparison of number symbols in adulthood (there is both supporting and contradicting evidence for these suggestions, as will be discussed in Chapter 2).

In both sentences, the one with 'many rabbits' and the one with 'five rabbits', my friend has used a symbol, the word 'many' or 'five', to refer to a quantity that they have just observed perceptually. Therefore, both 'many' and 'five' can be seen as symbols referring to stored quantity information. The difference between the meanings of 'many' and 'five' is that the latter refers to a precise quantity — I know exactly how many rabbits my friend has seen. The number symbol provides a sharp distinction between numerical magnitudes whereas the quantifier 'many' does not have a precise quantity reference. Many quantifiers refer to approximate quantities, making them compatible with the way we perceive and represent perceptually assessed quantities. On the other hand, number symbols that refer to precise quantities cannot in fact be directly related to perceptually

given quantities because these perceptually given quantities are only approximate. For this reason, I suggest that natural language quantifiers are even more likely to interact with and rely on neurocognitive systems for processing nonsymbolic quantity than are number symbols. The idea that the processing of at least some quantifiers may involve representations and processing mechanisms that are used for perceptually given quantities has been suggested before, but not discussed extensively (e.g., Clark & Grossman, 2007; Lidz, Pietroski, Halberda, & Hunter, 2011; Pietroski, Lidz, Hunter, & Halberda, 2009; Solt, 2011).

If we want to understand how and whether quantifier processing interacts with, or partially relies on, the cognitive mechanisms for estimating and comparing quantities from perceptual input, it may be useful to consider quantifier processing from the perspective of what is already known about number symbols. In Chapter 2, we try to build this bridge between perceptually given quantities and linguistic quantifiers. Chapter 2 thus presents an extensive review of what is known about number symbol processing and, based on this review, identifies possible directions of research and paradigms that can be applied to the investigation of quantifier processing.

1.3 Scalar adjectives as symbolic references to magnitude information

In this thesis, I assume that scalar adjectives can also be seen as symbolic references to magnitude information. Whereas quantifiers and number symbols refer to quantity information, scalar adjectives refer to magnitudes along other dimensions such size, length, width, duration, etc. We know that humans are also capable of estimating and comparing magnitudes along these perceptual dimensions.

The way humans perceive and compare magnitudes in these perceptual dimensions is considered to be parallel to quantity perception - the estimates are noisy and performance in comparison depends on the ratio between two magnitudes rather than absolute magnitudes (Cohen Kadosh, Lammertyn, & Izard, 2008). Again, this ability is considered innate, as infants and other animals have it as well (Feigenson, 2007; Tudusciuc & Nieder, 2009; Vallentin & Nieder, 2008). In fact, it has been suggested that a single shared mechanism exists that is involved in the perception and in the comparison of magnitudes along various dimensions (including numerical magnitude): the *generalized magnitude representation system* (Gallistel & Gelman, 2000; Lourenco, Ayzenberg, & Lyu, 2016; Walsh, 2003). I have already discussed how the properties of vague quantifier meanings suggest that they may interact with or partially rely on cognitive mechanisms for perceptually assessed numerical magnitudes. Similarly, in Chapter 4 we suggest that scalar adjectives may interact with or partially rely on cognitive mechanisms for perceptually assessed magnitudes in the generalized magnitude representation

system. For scalar adjectives, we conducted a series of behavioral experiments that test a specific hypothesis: that retrieval of the meaning of scalar adjectives requires the involvement of the generalized magnitude representation system.

In line with the approach taken here, some explanations in formal semantics have proposed that scalar adjectives are vague because they refer to noisy internal magnitude representations. One such explanation of the vagueness of scalar adjectives that takes human magnitude perception into account comes from Wright (Wright, 1975; see also e.g., Égré & Barberousse, 2014 for a discussion of the connection between scalar adjectives and approximate magnitudes suggested by Borel as early as 1907). In his discussion of scalar adjectives, Wright emphasizes the inability of our perceptual system to distinguish two objects with respect to the relevant property based on simple observation alone. Specifically, when the change in the degree of magnitude (e.g., of height, length, size, duration, etc.) is too small to be perceived (e.g., the difference between 190 cm and 189 cm which I brought up earlier), then it cannot affect our judgment about the applicability of a scalar adjective. Thus, the difference between magnitudes needs to be perceptually clearly identifiable in order to affect our use of scalar adjectives (a similar suggestion is that there needs to be a *gap* between the magnitudes; Pagin, 2010; van Rooij, 2011a). In a recent model of judgments about the applicability of scalar adjectives (Égré, 2017; see also Fults, 2011 for a proposal along similar lines), Egré proposes that in deciding whether a scalar adjective is applicable as a description of an object, the object's magnitude (regardless of whether it has been observed perceptually or described with exact numbers) first needs to be mapped to an inner scale of magnitude representation that makes this magnitude necessarily imprecise and approximate. I do not discuss this or other models in detail here, as we did not investigate the particular way in which magnitude representations come into play. Instead, the experiments reported in Chapter 4 test whether such magnitude representations are recruited in the processing of the meaning of scalar adjectives.

1.4 The context-sensitivity of scalar adjectives reflected in language composition at the neuronal level

As discussed above, the meaning of scalar adjectives is to a large extent context-sensitive. More specifically, the meaning of a scalar adjective depends crucially on the noun with which the adjective is combined. In contrast, the meaning of non-gradable adjectives does not depend on context to the same degree. It is assumed that whereas in the case of scalar adjectives a comparison class needs to be determined and a threshold needs to be computed, there is no need to determine a comparison class and compute a threshold for non-gradable adjectives.

This difference should in principle be reflected in the neuronal processes computing the meaning of an adjective–noun composition, i.e., for the way how our brain combines the meaning of an adjective and a noun to an integrated meaning representation of the noun phrase. Such a difference in the neuronal correlates of the processing of adjective-noun-phrases with scalar adjectives as opposed to non-gradable adjectives has been observed in a recent study by Ziegler and Pylkkänen (2016). This finding is intriguing because it goes beyond the level of formal semantics and provides support for a difference in processing different adjective classes at the neuronal level. In addition, it opens up the possibility of a number of follow-up studies that investigate composition mechanisms for scalar and other adjectives more closely.

Compositionality, i.e., the combination of multiple units of meaning into a new, integrated meaning, is one of the fundamental properties of human language and is a subject of a long tradition of empirical research in psycholinguistics and cognitive neuroscience of language. One empirical approach has used magnetoencephalography (MEG) data measured during the processing of minimal two-word phrases (Bemis & Pylkkänen, 2011, 2013a, 2013b). This is also the approach taken by Ziegler and Pylkkänen (2016). In their experimental paradigm, participants saw an adjective followed by a noun and subsequently responded to a question about the phrase. They contrasted the processing of adjective-noun-phrases with scalar adjectives versus non-gradable[4] adjectives. They observed a difference in the level of neural activity in the left anterior temporal lobe (LATL). Based on this result, Ziegler and Pylkkänen suggest that the composition of a noun with a scalar adjective happens at a later point in time than the composition of a noun with a non-gradable adjective. This difference is thought to arise due to the need to compute a context-sensitive threshold based on the noun's meaning in case it is combined with a scalar adjective. This process is not needed in the case of an adjective-noun phrase with a non-gradable adjective.

Considering the potentially far-reaching implications of the findings reported by Ziegler and Pylkkänen (2016), the study presented in Chapter 5 was conducted to ensure that the effect is robust. Specifically, we wanted to see whether we were able to observe the difference between the processing of noun phrases with scalar adjectives versus non-gradable adjectives in our own set-up and in a different language (Dutch). Beyond this replication attempt, the design of our study also should allow to tease apart semantic composition processes and syntactic composition processes, two types of composition processes that could not be distinguished in previous research.

[4]Note that in the corresponding chapter, we refer to these as *intersective* adjectives, in keeping with the terminology used by Ziegler and Pylkkänen.

1.5 Methodological remarks

Over the past decade, and increasingly throughout the years that I have spent conducting the research reported in this thesis, various issues with the robustness of findings reported in experimental psychology and cognitive neuroscience have been brought forward. Specifically, due to factors like low sample sizes, publication bias, and similar issues, it has become apparent that a large proportion (Szucs & Ioannidis, 2017 estimate that it possibly exceeds 50%) of reported effects in psychology and cognitive neuroscience are likely false positives or are of negligibly small size (see e.g., Fanelli, 2012; C. J. Ferguson & Brannick, 2012; Ioannidis, 2008; Szucs & Ioannidis, 2017). This has been confirmed by a low rate of successful replications in several large-scale replication projects for claims in psychology (Camerer et al., 2018; Collaboration, 2015; R. A. Klein et al., 2018). Research into language processing is no exception, with several failed replications reported (Kochari & Flecken, 2019; Nieuwland et al., 2018; Vasishth, Mertzen, Jäger, & Gelman, 2018). Conducting a replication study is one way to gain more confidence that a particular effect is real and to get a more accurate estimate of the effect size (Schmidt, 2009; Zwaan, Etz, Lucas, & Donnellan, 2018). Given these considerations, in my own research I adopted a strategy where I first try to replicate the effect that I am attempting to build on before I collect data in planned follow-up research (Kochari & Ostarek, 2018). Adoption of this approach is behind two of the chapters presented in this thesis (Chapters 3 and 5).

There are several arguments for conducting a replication of a study that one is planning to build on in a new research project.[5] One might undertake a conceptual replication, where the effect is investigated in a similar, but not identical experimental set-up, or a close replication, where as many variables as possible are kept identical to the original study. The first reason to conduct a replication is, as mentioned, avoiding basing a new study on a false positive or on an effect that is too small to be of interest. Especially in cases of a different lab set-up, different target population, or different language, it is important to establish that one is able to reproduce the original effect. Second, the results of a replication study, either showing the same or a different effect as in the original study, are in any case useful to the scientific community. A successful replication will strengthen the confidence in the effect and will help to establish the true effect size of the effect, whereas a failed replication will prompt discussion of the reasons for the failure (Schmidt, 2009; Zwaan, Etz, et al., 2018).

If a follow-up study shows no effect or supports the null hypothesis after one has successfully replicated the original effect, it becomes easier to interpret the the outcome of the follow-up study because it can be directly contrasted with the original (replicated) effect. Similarly, a successful replication published along

[5]Part of the text in this paragraph was previously published in an opinion piece, Kochari & Ostarek, 2018.

with the effect in the follow-up study make for a solid package that will convince the researchers themselves as well as colleagues who read their work. Finally, given that any outcome (including not observing any effect) can be meaningfully interpreted, the replication-first approach shifts the focus from the significance of the results (which is what is currently rewarded in psychology, and which is at least partially responsible for the current robustness issues; Gelman, 2018; Ioannidis, 2008; Vasishth, Mertzen, et al., 2018) to the methodological rigor of the research project and to the sizes of the observed effects.

Adopting this approach, I devoted considerable time to the replication of the studies on which I planned to base my own investigations. In one case (Chapter 3) the replication was successful and allowed us to proceed with my own experiments (presented in Chapter 4). In a second case (Chapter 5) we did not replicate the effect on which we had planned to build in a follow-up study (see the *Summary and avenues for future research* chapter for more on this). The failure to observe an effect in this case prompted us to conduct a thorough investigation of the set-up of the studies that reported such an effect in the past and to formulate a number of new hypotheses about the potential modulating factors of the effect.

One of the ways to enhance the robustness and replicability of experimental results is to collect data with larger sample sizes and a more diverse set of participants. Web-based data collection allows one to do just that and has several additional advantages over traditional lab-based studies. Specifically, it allows for faster data collection, geographical flexibility, and lower cost of administration. Furthermore, experiments created for web-browsers can be more easily shared between researchers. For all of these reasons, the behavioral experiments presented in Chapter 4 of this thesis were conducted remotely, in the web-browsers of participants. Ensuring that web-based data collection is technologically feasible and will result in comparable data quality to physical lab-based data was another motivation for the work reported in Chapter 3 of this thesis. There, I explore the suitability of two experimental paradigms from the numerical cognition literature for testing the hypothesis about scalar adjectives. Because both of these paradigms are influential in numerical cognition research and have been used in a large number of studies in the past, I decided to write a separate paper (included here as Chapter 3) where I report my experience with web-based data collection and the results of the replications. This chapter is intended as a contribution to the field of numerical cognition research, but the more general introduction and discussion would be useful to any researcher starting with web-based data collection.

1.6 Overview of manuscripts corresponding to chapters

Contributions defined according to CRediT, https://casrai.org/credit/.

Chapter 2:
Kochari, A., & Szymanik, J. Questions about Quantifiers: symbolic and nonsymbolic quantity processing by the brain. *Manuscript.*

> Conceptualization, Methodology: A.K., J.S.; Investigation, Project administration, Writing – original draft: A.K.; Supervision - J.S.; Writing – review & editing: A.K., J.S.

Chapter 3:
Kochari, A. (2019). Conducting web-based experiments for numerical cognition research. *Journal of Cognition, 2*(1): 39, 1-21. doi: 10.5334/joc.85.

Chapter 4:
Kochari, A., & Schriefers, H. Processing symbolic magnitude information conveyed by number words and scalar adjectives: parallel size congruity and same/different experiments. *Manuscript*

> Conceptualization, Methodology: A.K., H.S.; Data curation, Formal analysis, Investigation, Project administration, Writing – original draft: A.K.; Supervision - H.S.; Writing – review & editing: A.K., H.S.

Chapter 5:
Kochari, A., Lewis, A., Schoffelen, J.-M., & Schriefers, H. Semantic and syntactic composition of minimal adjective-noun phrases in Dutch: An MEG study. *Manuscript.*

> Conceptualization, Methodology: A.K., A.L., H.S.; Data curation, Investigation, Project administration, Writing – original draft: A.K.; Formal analysis, Visualization: A.K., A.L., J-M. S.; Writing – review & editing: A.K., A.L., H.S., J-M. S.

Chapter 2

Questions about Quantifiers: symbolic and nonsymbolic quantity processing by the brain

Abstract

[1]One approach to understanding how the human cognitive system stores and operates with quantifiers such as 'some,' 'many,' and 'all' is to investigate their interaction with the cognitive mechanisms for estimating and comparing quantities from perceptual input (i.e., nonsymbolic quantities). While a potential link between quantifier processing and nonsymbolic quantity processing has been considered in the past, it has never been discussed extensively. Simultaneously, there is a long line of research within the field of numerical cognition on the relationship between processing exact number symbols (such as '3' or 'three') and nonsymbolic quantity. This accumulated knowledge can potentially be harvested for research on quantifiers since quantifiers and number symbols are two different ways of referring to quantity information symbolically. The goal of the present review is twofold. First, we provide an overview of findings and methods from research into the relationship between number symbol and nonsymbolic quantity processing that could be of potential use for understanding quantifier processing. Research from the developmental, behavioral, and neuronal perspectives is reviewed. Second, we present an extended discussion of the properties of various quantifier classes in relation to the properties of nonsymbolic quantity processing mechanisms and research conducted about this relationship so far. Importantly, while doing so, we also provide a set of research directions and specific questions for the investigation of quantifier processing, in parallel with and

[1]This chapter is based on: Kochari, A., & Szymanik, J. Questions about Quantifiers: symbolic and nonsymbolic quantity processing by the brain. *Manuscript*.

inspired by the overview of research about number symbols.

2.1 Introduction

Humans can perceive, represent, and compare perceptually extracted quantities, e.g., extracted from visually presented arrays of objects or from aurally presented series of tones, as well as quantities that are presented using arbitrary symbols and natural language. In the former case, we can make an approximation of the quantity of elements (i.e., the *cardinality* or *numerosity*). In the latter case, a set of conventions can be learned to represent the exact cardinality using number symbols (e.g., Arabic digits, number words, Roman numerals — 7, 'seven', VII). In addition, approximate cardinality and the relationship between cardinalities can be conveyed symbolically using natural language quantifiers (e.g., 'some', 'many', 'most' etc.). While a lot of research has been devoted to the question of how number symbols and nonsymbolic quantities (we refer to perceptually extracted quantities as *nonsymbolic*) are linked in the human brain, substantially less is known about the link between quantifiers and nonsymbolic quantities. In this paper, we give a comprehensive review of experimental research on the relation between number symbols and nonsymbolic quantities and relate it to parallel research between quantifiers and nonsymbolic quantities. Most importantly, we put forward a set of new questions about the neurocognitive underpinnings of quantifier semantics. We hope that these questions not only influence research directions of experimental semantics but also help in building a bridge between number cognition and formal semantics communities.

The processing of number symbols has been the subject of extensive research within numerical cognition, given that number symbols referring to exact quantities play an important role in everyday functioning in modern industrialized cultures and are used in mathematics (e.g., Eger, 2016; Nieder, 2016; Sokolowski & Ansari, 2016). As the nonsymbolic quantity representation system is considered to be evolutionarily old and innate in humans, particular attention has been paid to the interaction of number symbols with, and their possible reliance on, the nonsymbolic quantity processing mechanisms. In the first half of the paper, we provide an up-to-date comprehensive review of developmental, behavioral, and neuronal-level evidence accumulated in the symbolic and nonsymbolic quantity processing research in relation to number symbols. This review functions as a backdrop for our discussion of parallel questions in relation to natural language quantifiers.

Natural language quantifiers are pervasive in everyday communication and are used as tools to refer to quantity even in individuals and cultures without extensive exact number symbol systems. Natural language quantifiers also potentially interact with and rely on neurocognitive systems for processing nonsymbolic quantity. Suggestions about the existence of a potential link between quantifiers

and nonsymbolic quantity processing have been put forward before (e.g., Clark & Grossman, 2007; Coventry, Cangelosi, Newstead, Bacon, & Rajapakse, 2005; Holyoak & Glass, 1978; Pietroski et al., 2009), but never discussed extensively. In the second half of the paper, we offer an extensive review of the relatively few published studies looking at this relationship. We then suggest directions for future research in this line by formulating a set of questions. Our goal here is to use existing research questions into number symbols and paradigms used in this regard to help formulate new questions about quantifiers. We believe that enriching the research on quantifier processing by taking accumulated knowledge regarding number symbols into account is a fruitful way forward. The reader should bear in mind that the goal of this manuscript is to start or stimulate discussion, and, hence, some ideas in which the details are not fleshed out and some speculative suggestions are also presented.

2.2 Nonsymbolic quantities and number symbol processing

2.2.0.1 Nonsymbolic quantities

Adults with and without formal education (see e.g., Barth, Kanwisher, & Spelke, 2003; Barth, La Mont, Lipton, & Spelke, 2005; Ferrigno, Jara-Ettinger, Piantadosi, & Cantlon, 2017; Gordon, 2004; Pica et al., 2004), pre-linguistic infants (see e.g., Izard et al., 2008; Wynn, 1992; Xu & Spelke, 2000), and both trained and untrained animals (see e.g., Brannon & Terrace, 1998; Breukelaar & Dalrymple-Alford, 1998; Cantlon & Brannon, 2006, 2007; Scarf, Hayne, & Colombo, 2011) are all able to approximate the quantity of items in sets and compare them in terms of their cardinalities (i.e., the total quantity of items in a set) when presented visually or aurally. In all cases, when comparing the cardinalities of sets, performance in terms of accuracy and reaction times has been observed to depend on the ratio between cardinalities rather than the absolute quantity of items in each set: the larger the ratio between quantities, the better the performance. For example, when asked to choose a set with a larger cardinality among sets of 3 and 5 items, responses are given faster and they are more accurate than when choosing among sets of 7 and 9 items (same for 30 and 50 vs. 70 and 90 etc.). Children and animals need higher ratios between cardinalities of sets to be able to successfully distinguish between them than human adults do, but they also exhibit ratio-dependent performance. This common pattern suggests that the mechanism for approximation of cardinalities might have the same evolutionary origin across the species.[2]

[2]Another piece of evidence for the approximate cardinality representation mechanism having the same evolutionary origin in humans and other animals comes from the observation of number-sensitive neurons in homologue areas of human and monkey brains (e.g., Harvey, Ferri,

The ratio-dependent performance indicates that there is a certain level of uncertainty (or noise) in underlying psychological representations of the quantities of nonsymbolically presented sets, and that the amount of uncertainty is proportional to the value, with larger numerical values having more noise. Values closer to one another will have a larger overlap in their representations than values further away from one another (e.g., the uncertainty around 7 means that it overlaps with 6 and 8, but less so with 5 and 9, etc.). Such ratio-based performance follows Weber's law for the perception of continuous stimulus dimensions. There is discussion around the exact way in which the nonsymbolic quantities are represented — linearly with scalar variability or logarithmically with fixed variability (e.g., Bar, Fischer, & Algom, 2019; Dehaene, Izard, Spelke, & Pica, 2008; Feigenson et al., 2004; Merten & Nieder, 2008), but we leave this discussion aside as it is not relevant for the purposes of the present paper. This nonsymbolic quantity representation mechanism, at least in case of quantities above 4 (see next paragraph), is often referred to as the Approximate Number System (ANS; but see Núñez, 2017 for a discussion of the terminology in this research line).

The only exception to ratio-based performance with nonsymbolically presented quantities is the case of quantities up to 3 or 4. Behavioral performance with these quantities does not seem to be ratio-dependent, which suggests that they might be represented without uncertainty (i.e., exactly).[3] The process by which these quantities are estimated or extracted is traditionally referred to as *subitizing*. Some have explained this difference by suggesting that small-magnitude numbers are represented with very little noise within an ANS-like system (e.g., Cordes, Gelman, Gallistel, & Whalen, 2001; Dehaene, 2007), whereas others have posited that a separate mechanism represents them exactly — the *object tracking system* (also referred to as *parallel individuation system*; e.g., Carey, 2009; Cordes et al., 2001; Feigenson et al., 2004; Hutchison, Ansari, Zheng, Jesus, & Lyons, 2019; Kaufman, Lord, Reese, & Volkmann, 1949; Revkin, Piazza, Izard, Cohen, & Dehaene, 2008; Trick & Pylyshyn, 1994). The object tracking system is also considered to be innate and might play a role in number symbol learning (Carey, 2001, 2009). While subitizing is relevant in multiple discussions in number symbol processing research, in the present paper we only discuss it in the context of number symbol learning since, as we see it, only in that case can it subsequently be related to research on quantifiers.

& Orban, 2017; Nieder & Dehaene, 2009; see also Ferrigno, Hughes, & Cantlon, 2016).

[3]Additional evidence for a separate mechanism being responsible for small quantities comes from studies on eye-movement patterns (Watson, Maylor, & Bruce, 2007) and visual working memory (Piazza, Fumarola, Chinello, & Melcher, 2011).

2.2.0.2 Number symbols

Humans can refer to exact cardinalities using formal symbolic systems such as, e.g., number words, Arabic digits, Japanese Kanji, Roman numerals, etc.[4] Number symbols differ from nonsymbolically presented cardinalities in that they refer to exact quantities, require a culture to have developed such a formal system, and at least larger numbers need to be explicitly taught to children (Pica et al., 2004). It has been proposed that number symbol representations and processing mechanisms are based on nonsymbolic quantity representations and processing mechanisms evolutionarily, developmentally, and in terms of neuronal implementation (e.g., Dehaene, 1997, 2007; Dehaene & Cohen, 2007; Feigenson et al., 2004; Nieder, 2016), though this view has been a subject to criticism and counter-evidence has been reported in recent years (e.g., Carey & Barner, 2019; Carey, Shusterman, Haward, & Distefano, 2017; Krajcsi, Lengyel, & Kojouharova, 2016; Núñez, 2017; Reynvoet & Sasanguie, 2016).

One influential hypothesis is that number symbols, as a novel cultural invention, invade evolutionarily older brain circuits responsible for nonsymbolic cardinality processing — the so-called 'neuronal recycling hypothesis' (Dehaene & Cohen, 2007). The proposal is that since number symbols have not been around long enough for dedicated neuronal machinery to have evolved, this new function has to be embedded in circuits that have originally evolved for something else. On a strong version of this hypothesis, as a consequence of this 'recycling', the same neuronal populations should represent number symbols and nonsymbolic cardinalities (e.g., Dehaene, 2007; Gallistel & Gelman, 1992). On a weaker version of the hypothesis, some representational properties of number symbols are inherited from nonsymbolic cardinality representations and/or information processing mechanisms (e.g., working memory storage, attentional, comparison and calculation mechanisms) should be shared between them (Lyons & Beilock, 2018; Nieder, 2016). It could be the case, for example, that number symbol representations rely on nonsymbolic cardinalities early in childhood when children begin to relate number symbols to quantities, but over the course of development number symbol and nonsymbolic quantity representations become distinct from each other and, therefore, come to refer only to exact cardinality representations (Lyons, Ansari, & Beilock, 2012; Matejko & Ansari, 2016). On an alternative hypothesis, number symbol representations are not based on nonsymbolic quantities at all but form a separate system of exact symbol–symbol relations (e.g., Krajcsi et al., 2016; Lyons & Beilock, 2018; Reynvoet & Sasanguie, 2016; Sasanguie, De Smedt, & Reynvoet, 2017), and are learned, e.g., by being associated with the object tracking system (Carey, 2001, 2009). In the rest of this section,

[4]While there are certain differences in their processing (e.g., number words need to be processed as words, which require phonological processing, whereas Arabic digits are single symbols which may not require phonological processing), here we omit discussion of these differences. Instead, we are interested in the higher-level, quantity representations they evoke.

we review obtained evidence and discussions regarding the proposals described in this paragraph.

This review is mostly focused on evidence and experimental paradigms relevant to our subsequent discussion on quantifiers (for more comprehensive reviews of this research and methodological discussions the reader is referred to e.g., Carey & Barner, 2019; Eger, 2016; Nieder, 2016; Nieder & Dehaene, 2009; Núñez, 2017; Sokolowski & Ansari, 2016; Wilkey & Ansari, 2019).

2.2.1 The developmental perspective

There is a lot of evidence suggesting that the nonsymbolic quantity processing system is innate in humans (Izard et al., 2008; Spelke, 2011; Wynn, 1998; Xu & Spelke, 2000). On the other hand, number symbols are learned during the lifetime of a person through explicit training. What do these number symbols referring to/grounded on? This is the symbol-grounding problem (Harnad, 1990) for number symbol representations. Here, we briefly discuss two main theoretical suggestions in the literature though neither can account for the whole process of number symbol acquisition without assuming additional mechanisms (see Carey & Barner, 2019; Carey et al., 2017, for a discussion).

2.2.1.1 Number symbols are grounded in approximate nonsymbolic representations

Given that the approximate quantity representation mechanism seems to be an evolutionarily old mechanism present in humans from birth, it has been suggested that ANS representations provide a basis for learning number symbols (Dehaene, 2007; Dehaene & Cohen, 2007; Feigenson et al., 2004; Gallistel & Gelman, 1992; Halberda, Mazzocco, & Feigenson, 2008). On a strong version of this hypothesis, "when we learn number symbols, we simply learn to attach their arbitrary shapes to the relevant nonsymbolic quantity representations [...] thus, the symbol '3' comes to evoke the very same representation that would be evoked by a set of three dots" (Dehaene, 2007; Gallistel & Gelman, 1992). This hypothesis is most theoretically parsimonious since it suggests that symbolic quantity simply makes use of a quantity mechanism already present in humans. Traditionally, findings of similar behavioral effects for processing both nonsymbolic and symbolic quantities in adults and of overlapping brain regions responsible for them have been thought to support this hypothesis, though there is substantial debate about the strength of the evidence there (this evidence is discussed in detail below, in sections 2.2.2 and 2.2.3). Moreover, these findings of similar effects are reported with adult participants, so they could be explained by the fact that number symbols and nonsymbolic quantities become associated with each other later in life rather than through the acquisition process.

Let us consider evidence supporting this hypothesis from studies on the language acquisition process. Some studies show an association between children's approximate quantity perception on the one hand and their understanding of number words and counting using the number symbol sequences on the other hand (Mussolin, Nys, Leybaert, & Content, 2012; Odic, Le Corre, & Halberda, 2015; Wagner & Johnson, 2011), but this has not always been observed (Carey et al., 2017; Le Corre & Carey, 2007). There is also evidence for children with a specific math learning disorder (dyscalculia) having less precise approximate number representations than their typically developing peers (e.g., Mazzocco, Feigenson, & Halberda, 2011; Piazza et al., 2010). Another line of research reports a correlation between performance in approximate quantity comparison tasks and symbolic mathematics achievement, suggesting that performing symbolic mathematics tasks may rely on approximate nonsymbolic quantity representations (e.g., Halberda, Mazzocco, & Feigenson, 2008; Keller & Libertus, 2015; Libertus, Feigenson, & Halberda, 2011; Mundy & Gilmore, 2009). Moreover, in some studies children trained in nonsymbolic approximate quantity comparison tasks show improved performance in symbolic math tasks (e.g., D. C. Hyde, Khanum, & Spelke, 2014; see also Park & Brannon, 2013 for similar results with adults). However, again a number of studies fail to find a correlation between nonsymbolic quantity processing and symbolic math achievement (see De Smedt, Noël, Gilmore, & Ansari, 2013; Reynvoet & Sasanguie, 2016, for reviews). Meta-analyses that looked at the combined evidence from these studies found that the association between nonsymbolic quantity processing and math achievement is small (Chen & Li, 2014; Fazio, Bailey, Thompson, & Siegler, 2014; Schneider et al., 2017), suggesting that nonsymbolic quantity representations probably contribute rather little to symbolic quantity processing. The validity and consequences of these and other findings from correlational and training studies are a subject to a lot of debate, which we skip here (see, e.g., Carey et al., 2017; Leibovich & Ansari, 2016; Reynvoet & Sasanguie, 2016, for reviews and discussion).

2.2.1.2 Number symbols are derived from the object tracking system

An alternative hypothesis about the development of number symbols posits that they develop completely independently of nonsymbolic quantity representations. We start by describing two observations on which this hypothesis is founded. The first is that there seems to be a certain order in which children learn number symbols (Piantadosi, Jara-Ettinger, & Gibson, 2014; Wynn, 1992). Specifically, children seem to learn the quantity that number words 'one,' 'two,' and 'three' refer to consecutively, taking some time before understanding the next number word. At the point when children learn the meaning of 'four,' they understand the principle of counting (i.e., *the cardinality principle*) — that the last number word when counting refers to the total number of objects in a set. At this point, children immediately understand the quantity reference of number words beyond

four. The second relevant observation is that infants and adults seem to be able to represent nonsymbolic quantities up to 3 or 4 in an exact, discrete, manner, unlike the approximate analog representations of larger quantities. As already discussed, these are quantities within the subitizing range.

According to the hypothesis put forward by Carey and colleagues (Carey, 2001, 2009; Le Corre & Carey, 2007; see also Piantadosi, Tenenbaum, & Goodman, 2012), children first learn the sequence of number words ('one, two, three, four, five...') as a list of meaningless words — without ascribing a numerical sense to them. We know that children can learn such lists and even simulate counting without understanding that numbers refer to the quantity of objects (Wynn, 1990, 1992). Simultaneously, the singular and plural distinction (grammatical), as well as quantifiers and articles that children hear in language, facilitate their reasoning about quantities and number words. Specifically, for example, the meaning of the word 'one' is learned as a quantifier within natural language, given that it is used in everyday language more frequently than in counting and given that in some languages it is synonymous to a singular determiner like 'a'. This way, 'one' is first learned as a quantifier denoting a singular entity. Gradually, children learn to associate the number words 'one' through 'three' or 'four' to the corresponding quantity that is tracked by the object tracking system. At this point they do not yet understand that quantities can be generalized to the rest of the counting list. Finally, children learn that every next number refers to one more object than the previous one and can generalize this knowledge to the rest of the number words (see Carey, 2001, 2009, for details and references).

One set of results supporting this hypothesis shows that children who understand the quantity reference of number words only up to 'three' could be taught to associate 'four' with sets of four objects, but at the same time could not be taught to associate 'ten' with sets of ten objects (Carey et al., 2017). This has been interpreted as evidence for that the children at this point did not yet understand that number words can refer to quantities beyond four. That is a predicted stage of development if number symbol knowledge initially relies on the object tracking system, but is unexpected if number symbol knowledge results from mapping number words to approximate nonsymbolic quantities (on this latter hypothesis, there should be no difference between quantities up to and beyond four). Also supporting this hypothesis, a recent neuroimaging study with adults found more similarity between the neuronal correlates of processing symbolic and nonsymbolic quantities within the subitizing range than outside of this range (Lyons & Beilock, 2018; see also Hutchison et al., 2019 for behavioral evidence for stronger association within the subitizing range than outside).

However, contradicting the suggestion that there is no mapping at all between nonsymbolic quantity and symbolic number words, two to five-year-old children do seem to be able to make estimates of quantities when asked to perform an action a certain number of times. Specifically, when given instructions with number words to put a certain number of objects in a bowl or to tap a certain number

of times, children gave/produced approximately the required quantity, even for quantities greater than 4 (Gunderson, Spaepen, & Levine, 2015; Odic et al., 2015). These results suggest that there exists some sort of mapping, albeit only unidirectional. Another problem with the proposal is that one of its elements does not seem to hold up against empirical evidence — namely, quantifier knowledge and the singular–plural distinction is not clearly related to number symbol knowledge. One study reports that children's level of knowledge of number symbols is not correlated with their level of knowledge of natural language quantifiers, but instead correlates with age (Dolscheid, Winter, Ostrowski, & Penke, 2017, though see Barner, Chow, & Yang, 2009, where such correlation was present). Moreover, some children learning English interpret the article 'a' as approximate, allowing it to refer to one or two objects, while at the same time interpreting 'one' exactly; this speaks against the singular-plural distinction being behind learning 'one,' or 'one' being initially learned as a synonym of 'a' (Barner, Chow, & Yang, 2009; see also Barner, 2012). Without this component, it is not clear why the number symbols have to be learned consecutively given that the object tracking system by definition gives children access to all three or four quantities at the same age[5] (Starkey & Cooper, 1995).

2.2.1.3 Development later in life

So far, we have discussed how number symbols are grounded in existing mechanisms when children learn them for the first time. Number symbol representation and processing mechanisms in adults might either stay connected to these initial representations to some extent or develop independently. At least when a person is exposed to life in a society that makes extensive use of number symbols and/or the person is taught formal math systems, these initial number symbol representations have to change over the course of development.

Specifically, because the symbolic quantity has to be exact, representations for number symbols cannot be exactly the same as those for approximate quantities (Carey & Barner, 2019; Núñez, 2017 provide detailed arguments for this view). In this sense, it is unreasonable to expect exactly the same representations for nonsymbolic quantities and number symbols, as has been suggested by some theories. Instead, it is more likely that when we start using number symbols more extensively, they develop into — at least to some extent — an independent system of quantity representation (as put forward in the so-called 'symbolic estrangement' hypothesis, see Lyons et al., 2012; see also Matejko & Ansari, 2016; Wilkey & Ansari, 2019). When it comes to the object tracking system, one important aspect is that it is not even capable of supporting a cardinality beyond 4. Thus,

[5]Note, however, that theories suggesting that number symbol learning is grounded on approximate nonsymbolic representations do not have an explanation for the specific order of the learning of number word meanings either, though there one could potentially appeal to the maturing of the ANS-like system.

in this case too we assume that the object tracking system is used to learn number symbols first, but at a later point children have to develop a symbolic number representation system beyond the initial state.

2.2.1.4 Interim summary

Overall, when reviewing the accumulated evidence, there is currently no compelling evidence for either the suggestion that number symbols are mapped onto the approximate nonsymbolic quantities or that their acquisition is based on information provided by the object tracking system. Both suggestions remain subjects of debate and more research is needed in both directions. Importantly, regardless of which system number words are initially mapped to, subsequently an at least partially independent symbolic number system has to develop. This is because approximate number representations are not capable of representing exact cardinality and because the object tracking system is only able to represent quantities beyond the subitizing range.

2.2.2 The behavioral perspective

In this section, we review the accumulated behavioral evidence investigating whether and to what extent the cognitive systems processing number symbols and nonsymbolic quantities are shared as well as paradigms that have been used so far. The first, basic question that needs to be asked about number symbols and nonsymbolic quantities is whether there at least exists an interface for mapping between the two types of quantity and to what extent such mapping happens automatically (i.e., without explicit instructions). The results of tasks based on estimation and congruity are reviewed in this context. While a useful starting point, evidence for the automaticity of mapping does not allow us to draw strong conclusions about shared or distinct representations and processing mechanisms for symbolic and nonsymbolic quantities. That is because such mapping can, in principle, simply be an association that emerges as a result of co-occurrence in the natural world rather than a fundamental link between processing mechanisms.

Numerical magnitude comparison and numerical matching paradigms have been used to investigate the similarity of representation formats of number symbols and nonsymbolic quantities. Specifically, this line of research has looked at the potential ratio-dependence of performance in these tasks. A similar behavioral performance for number symbols and nonsymbolic quantities has been suggested to arise from the similar representational format of the two (perhaps due to number symbols initially being derived from nonsymbolic quantity representations). Distinct patterns of behavioral effects, on the other hand, would speak to different formats of representations. For example, unlike nonsymbolic quantities, number symbols might be represented in a discrete format, without uncertainty (noise) in representations (e.g., Bar et al., 2019; Krajcsi et al., 2016;

Reynvoet & Sasanguie, 2016; Sasanguie et al., 2017). Some of the proposals put forward in this line ascribe any similarities between number symbols to similarities in their frequencies of occurrence and co-occurrence (Krajcsi et al., 2016; Lyons & Beilock, 2018; Verguts, Fias, & Stevens, 2005).

A further question is whether overlapping neuronal populations represent number symbols and nonsymbolic quantities; we discuss evidence from neuroimaging studies in the next section, but behavioral evidence from priming studies is also relevant for this question. Importantly, it has been demonstrated that the same neuronal populations may support symbolic and nonsymbolic quantities alike, even if they are represented in different formats (Dehaene, 2007; Verguts & Fias, 2004).

2.2.2.1 Mapping between number symbols and nonsymbolic quantities

We know that adults are capable of finding correspondences between nonsymbolic and symbolic quantities. To look at the mapping more closely, in estimation tasks participants are briefly (so as to prevent them from counting) presented with an array of objects and asked to give a number symbol to indicate the cardinality (e.g., Crollen, Castronovo, & Seron, 2011; Dehaene et al., 2008; Izard & Dehaene, 2008; Revkin et al., 2008; Whalen, Gallistel, & Gelman, 1999). People typically give precise estimates for numbers in the subitizing range and estimates increasingly further away from the true cardinality as the cardinalities get larger. An underestimation bias is observed with increasing cardinalities in this task: e.g., when presented with 80 dots, people tend to give a number symbol below 80 (e.g., Crollen et al., 2011; Izard & Dehaene, 2008; Krueger, 1982). The extent of underestimation increases with increasing nonsymbolic cardinalities, and differs individually (Crollen et al., 2011). Finally, giving participants a reference cardinality (e.g., presenting an array of objects and labeling it 'thirty') biases their subsequent judgments, meaning that they use it as an anchor (Izard & Dehaene, 2008). The fact that we can give a symbolic number label to a nonsymbolically presented quantity and vice versa speaks to the existence of at least an interface between these two representations of quantity.

A related question is whether symbolic and nonsymbolic quantity information is automatically activated and integrated when presented simultaneously, without explicit instructions to do so. In a series of studies, digits or letter strings were presented superimposed on dot arrays, and the participants' task was to simply decide whether they saw digits or letter strings (binary choice). In the digit trials, the dot arrays presented in the background either matched or mismatched the quantity represented by the digits. There was no quantity judgment in this task, so participants did not have to process either the symbolic or nonsymbolic quantities in order to perform it. Despite this, participants were more accurate and gave faster responses in trials where the value of the number symbols and the number

of dots matched than in the trials where they mismatched (A. S. Liu, Schunn, Fiez, & Libertus, 2015; see also R. Liu, Schunn, Fiez, & Libertus, 2018 for similar results[6]). These results tentatively suggest that either the nonsymbolic quantity is automatically converted to or activates a corresponding symbolic quantity or vice versa. We see this as a tentative conclusion because it is based on a single study, so more data is needed.

2.2.2.2 The representation format of number symbols

One paradigm traditionally used to investigate the representation format of symbolic and nonsymolic quantities is a magnitude comparison task in which either two digits/number words or two nonsymbolic quantities (e.g., arrays of dots) are presented side by side, and the participants' task is to choose the numerically larger/smaller one. Both in the nonsymbolic and symbolic comparison tasks, participants display ratio-dependent performance in terms of error rates and reaction time (e.g., Barth et al., 2005; Buckley & Gillman, 1974; Moyer & Landauer, 1967; Smets, Gebuis, & Reynvoet, 2013; Smets, Moors, & Reynvoet, 2016). Therefore, based on these results, number symbols are thought to be represented by noisy overlapping representations for values, just like nonsymbolic quantities. However, the suitability of this task for tapping into the representation format has been questioned. Specifically, the ratio-dependence of performance in a comparison task like this in the case of number symbols can instead be explained in part by the set-up itself, where solely decision-making process could give rise to the observed pattern (see Kojouharova & Krajcsi, 2018; Van Opstal, Gevers, De Moor, & Verguts, 2008; Verguts & Van Opstal, 2014), and in part by differences in the relative frequencies of different number symbols (see Krajcsi et al., 2016; Verguts et al., 2005).

A paradigm that avoids the above-described issues is a numerical matching task in which participants have to decide whether two sequentially presented symbolic number stimuli refer to the same quantity (participants respond with either 'same' or 'different'). When different numbers are presented (e.g., '5' and 'seven'), the performance also depends on the ratio between the two quantities (Defever, Sasanguie, Vandewaetere, & Reynvoet, 2012; Smets et al., 2013; Van Opstal & Verguts, 2011; Verguts & Van Opstal, 2005). However, this effect has not always been observed (Cohen, Warren, & Blanc-Goldhammer, 2013). In addition, a recent study employing audio-visual matching (instead of the visual presentation of two number quantities) did not reveal any distance effect in a condition where participants matched number words and digits (both symbolic), while such effects were obtained in a nonsymbolic matching condition and a mixed symbolic

[6]Though it should be noted that unfortunately in this study the participants performed an intensive nonsymbolic quantity estimation task before the relevant task, which possibly put them in the mode of estimating quantities, so a replication without this confound is required to ensure robustness of the effect.

and nonsymbolic matching condition (Sasanguie et al., 2017). This result rather supports different formats of representation for number symbols and nonsymbolic quantities. Overall, thus, the evidence from the numerical matching task remains mixed.

Yet another paradigm that has been used is subliminal or overt priming, where the so-called *priming distance effect* cannot be explained as a purely task-related decision-based effect either. In a typical numerical priming paradigm, participants are asked to compare a number to a standard (e.g., to decide whether each presented number is higher or lower than *5*) or to name a number aloud. Before the target number is visible, however, participants are subliminally or consciously presented with a prime number. The priming distance effect refers to the result that decision reaction times or naming latencies are slower when the prime number is closer to the target number (e.g., the target *four* being preceded by prime *three* as opposed to being preceded by prime *one*). The priming distance effect has been observed for various notations of number symbols (Arabic digits and number words — e.g., Koechlin, Naccache, Block, & Dehaene, 1999; Naccache & Dehaene, 2001b; Reynvoet & Brysbaert, 1999; Reynvoet, Brysbaert, & Fias, 2002) as well as for nonsymbolic stimuli (dot arrays — e.g., Defever, Sasanguie, Gebuis, & Reynvoet, 2011; Herrera & Macizo, 2008; Roggeman, Verguts, & Fias, 2007; Sasanguie, Defever, Van den Bussche, & Reynvoet, 2011). However, it has been observed that the exact pattern of the priming is different for nonsymbolic and symbolic quantities (weaker priming in the case of number symbols than nonsymbolic quantities), speaking against exactly the same representation format (see Herrera & Macizo, 2008; Roggeman et al., 2007); to be continued below.

2.2.2.3 Overlapping neuronal populations

The above-described priming paradigm has also been used to look at whether the same or different neuronal populations represent number symbols and nonsymbolic quantities. In this case, an array of dots was presented as a prime subliminally and an Arabic digit as a target, or vice versa. If the same neuronal population is activated for both, the priming distance effect should be observed across quantity types. In such a set-up, the priming distance effect has been observed when the primes are dot arrays and the targets are Arabic digits, but not when the primes are Arabic digits and the targets were dot arrays (Herrera & Macizo, 2008; Roggeman et al., 2007). This evidence speaks against fully overlapping representations for symbolic and nonsymbolic quantities because if that were the case we would expect priming in both directions. Instead, one explanation that has been suggested is that number symbols and nonsymbolic quantities are represented by the same neuronal populations, but in different formats (Herrera & Macizo, 2008; Roggeman et al., 2007). In the proposed neuronal architecture, both symbolic and nonsymbolic quantities are represented in such a way that they overlap with their neighboring quantities, but number symbols have substantially

less overlap with neighbors (i.e., are substantially less noisy) than nonsymbolic quantities (architecture suggested in a computational model by Verguts & Fias, 2004). This then leads to sharper, more distinct representations for number symbols, though not completely discrete. Therefore, activation of the cardinality of a number symbol also spreads to neighboring cardinalities, but is limited to the closest neighbors, whereas activation of the cardinality of a nonsymbolic array spreads to neighboring cardinalities more widely.

2.2.2.4 Interim summary

While the results of a number of prominent studies support a similar format of representations for number symbols and nonsymbolic quantities, these paradigms have either been shown to be flawed (in the case of numerical magnitude comparison tasks) or have produced mixed results (in the case of numerical matching tasks). Stronger evidence comes from studies that made use of priming paradigms. These studies suggest that there probably is some overlap in the cognitive systems representing number symbols and nonsymbolic quantities. The observed pattern of effects is best explained by assuming that number symbols and nonsymbolic quantities have a different representational format (specifically, different amounts of overlap with neighboring quantities) within these cognitive systems.

2.2.3 The neuronal perspective

From the perspective of neural implementation, research has investigated which populations of neurons represent symbolic and nonsymbolic quantities (including whether some neuronal populations exist that are responsible for representing a quantity regardless of the presentation format, e.g., three dots and '3'), and whether the representational format, i.e., the way a quantity is coded by these neuronal populations, is similar for both (for detailed reviews and discussions of findings with the approaches discussed here and others see e.g., Eger, 2016; Piazza & Eger, 2016; Sokolowski & Ansari, 2016; Wilkey & Ansari, 2019).

Given that behaviorally the discrimination performance with nonsymbolic quantities depends on the ratio between the two quantities, one prominent proposal for how nonsymbolic quantity is implemented neurally is that there exist populations of neurons coarsely tuned to a preferred quantity (Dehaene, 2007; Dehaene & Changeux, 1993; Nieder, 2016; Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004). 'Coarse tuning' refers to the idea that quantity-selective neurons should respond most strongly to their preferred quantity and show progressively declining activity levels in response to quantities that are further away from their preferred quantity. Note that on this proposal, while there exist single quantity-selective neurons, quantity is not represented by a single neuron, but rather by the activity of a population of differently tuned neurons (several population-coding models have been proposed, for details see Nieder, 2016 and the references

therein). If number symbols are represented in the same format as nonsymbolic quantities, we would expect to see a neuronal population following the coarse tuning principle when responding to quantity information for number symbols as well (Dehaene, 2007; Piazza, Pinel, Le Bihan, & Dehaene, 2007).

2.2.3.1 Quantity-selective neurons

Single cell recordings in monkeys viewing arrays of dots have indeed identified quantity-selective neurons within different subregions of the parietal and prefrontal cortex (Nieder, Freedman, & Miller, 2002; Nieder & Miller, 2003, 2004; see Nieder, 2016 for a review). As expected based on the coarse tuning suggestion, these neurons were most responsive to a preferred quantity and less so to other quantities in a gradual manner based on how far they were from the preferred quantity. The only study to date that has used single cell recordings with humans viewing dot arrays also successfully identified neurons with firing patterns corresponding to the neural tuning hypothesis (Kutter, Bostroem, Elger, Mormann, & Nieder, 2018). However, the recordings in this study were in the medial temporal lobe, which is not the region most consistently observed for quantity processing in fMRI studies (these are the parietal areas, see below). When it comes to number symbol processing, Kutter and colleagues observed distinct neurons responding to number symbols, with response profiles substantially less aligned with the gradual activity decrease dependent on the distance. These results with humans, in principle, suggest a differential encoding of symbolic and nonsymbolic quantities. However, given that Kutter and colleagues only recorded neurons in the medial temporal lobe, it is still possible that neurons responding to both symbolic and nonsymbolic quantities, and with a similar response profile, exist in the parietal areas, which in fact are considered to be crucial for quantity processing.

2.2.3.2 Importance of the parietal cortex

Some evidence regarding the brain areas responsible for quantity processing in humans comes from studies with patients with brain atrophy. Specifically, deficits in quantity processing have been reported for patients with Corticobasal Syndrome (CBS), which is associated with atrophy most prominently in the parietal cortex. These patients have difficulty comparing both symbolically or nonsymbolically presented quantities (e.g., Halpern et al., 2004; Koss et al., 2010; McMillan, Clark, Moore, & Grossman, 2006) and carrying out addition and subtraction operations even with small numbers (e.g., Halpern, McMillan, Moore, Dennis, & Grossman, 2003; Spotorno, McMillan, Powers, Clark, & Grossman, 2014). Simultaneously, they typically do not have impaired speech or problems understanding other concepts (e.g., they do not have a deficit in object naming). Therefore, the parietal cortex seems to house neuronal populations that play a crucial role in quantity processing. The fact that the deficit occurs for both symbolic and nonsymbolic

quantities suggests that both are housed in the parietal lobe.

The importance of the parietal cortex for quantity processing is also confirmed by numerous fMRI studies. In fMRI studies using various tasks, both symbolic and nonsymbolic number processing have been shown to involve regions of the prefrontal cortex and parietal cortex (for reviews, see Arsalidou & Taylor, 2011; Sokolowski & Ansari, 2016; Sokolowski, Fias, Mousa, & Ansari, 2017). Most consistently, activity in the intraparietal sulcus and areas around it has been correlated with processing Arabic digits, number words, dot arrays, etc.; in numerical magnitude comparison tasks, during arithmetic tasks, during passive viewing, etc. Moreover, the amount of activity in these regions has been observed to be sensitive to the exact quantity that is being processed, regardless of whether it is a number symbol or a nonsymbolic quantity (e.g., Eger et al., 2009; Lyons, Ansari, & Beilock, 2015; Lyons & Beilock, 2018; Piazza et al., 2007). However, given the limited spatial resolution of fMRI data, a similar average or total amount of activity within an area is not sufficient to conclude that the specific set of neurons that are involved in both cases is the same (a similar activation in the area could also arise from distinct neuronal populations housed close by in that area) or that the representational format was the same. For this reason, more advanced analysis methods have been used in recent years, to which we now turn.

2.2.3.3 Representation format and overlapping neuronal populations

One paradigm that has been used to look at the brain areas responsive to quantity processing with a better resolution is the fMRI adaptation paradigm (Grill-Spector, Henson, & Martin, 2006; Naccache & Dehaene, 2001a). The fMRI adaptation paradigm is based on the observation that with repeated presentations of the same stimuli, activation of the neurons that specifically represent this stimulus at an object level is reduced (this is referred to as 'adaptation'). When a different object (a 'deviant') is presented, the activation level of these neurons increases again. Importantly, these neurons are thought to represent the stimulus at an object level because the reduction in activity occurs even if other factors (such as size, color, location, etc.) change over repetitions; i.e., it is invariant to lower-level sensory changes. This allows researchers to identify regions coding for stimulus categories of interest and to probe specific represented features by manipulating the features of the deviants. In a number of fMRI studies, neuronal populations in the left and right intraparietal sulci and surrounding areas adapted to the cardinality of dot arrays and showed an increase in activation levels when presented with a novel cardinality (Cantlon, Brannon, Carter, & Pelphrey, 2006; He, Zhou, Zhou, He, & Chen, 2015; Piazza et al., 2004). The amount of activation increase in these studies was modulated by the distance between the adapted-to and deviant cardinalities, corresponding to what is expected under the coarse tuning hypothesis. Parallel adaptation and distance-dependent increase in activity in the intraparietal cortex in response to a deviant has also been observed for num-

ber symbols[7] (Goffin, Sokolowski, Slipenkyj, & Ansari, 2019; Holloway, Battista, Vogel, & Ansari, 2012; Notebaert, Nelis, & Reynvoet, 2010; Notebaert, Pesenti, & Reynvoet, 2010; Vogel et al., 2017). Importantly, in one study this effect was observed even when the participants adapted to nonsymbolically presented cardinality, whereas the deviant was a number symbol and vice versa (Piazza et al., 2007). These results make a strong case for the view that the same neuronal populations may be representing quantity in both notations.[8] Overall, the results with the adaptation paradigm support the possibility that the intraparietal sulcus houses representations for both symbolic and nonsymbolic quantities according to the coarse tuning hypothesis, possibly in overlapping neuronal populations that represent the quantity.[9]

Let us now take a quick detour to discuss other findings with the adaptation paradigm that are also relevant to our discussion with quantifiers below. Besides investigating cardinality representations, this paradigm has been used to investigate whether a ratio between two simultaneously presented cardinalities is coded, and whether the ratio representations are organized in an analog ANS-like system similar to approximate nonsymbolic quantity representations (i.e., with overlapping representations with neighboring ratio values). Indeed, the few studies that have been conducted to date suggest that, parallel to cardinality, ratio information also seems to be encoded in the intraparietal sulcus and exhibits distance-dependent neuronal activity recovery when presented both symbolically (Jacob & Nieder, 2009a) and nonsymbolically (Jacob & Nieder, 2009b; Jacob, Vallentin, & Nieder, 2012). Based on these results, it has been suggested that ratio is also coded with coarse tuning, and that both numerosity and ratio are processed by the same brain area.

Another series of studies used the representational similarity analysis (RSA; Kriegeskorte, Mur, & Bandettini, 2008) to investigate neuronal activation patterns for nonsymbolic quantities and number symbols. The idea behind RSA is to compare the similarity of activation patterns within a brain region of interest (rather than e.g., looking at the average level of activation in this region) in response to pairs of stimuli (or conditions) and, based on that, make an inference about the information represented in this region. Actual pattern similarity can be compared to predicted similarity by various models with the goal of identifying the model that best explains the neural activity patterns. It has been observed

[7] In fact, it has been proposed that specifically the left intraparietal cortex is more involved in representing symbolic number (Ansari, 2007; Sokolowski, Fias, Mousa, & Ansari, 2017), but we will omit the discussion of this point here.

[8] However, see also Cohen Kadosh et al., 2011, who observe an increase in intraparietal cortex activity following adaptation for a notation-only change (i.e., when the deviant is a non-symbolic array with the same quantity as the adapted-do number symbol) and argue for distinct neuronal populations representing symbolic and nonsymbolic cardinalities within this region.

[9] We say only *possibly* because to our knowledge only a single study to date has demonstrated cross-notation adaptation.

that similarity of activity patterns in intraparietal cortex (and other regions) when processing different nonsymbolic quantities is correlated with the numerical distance between these quantities, in line with the coarse tuning hypothesis (Lyons et al., 2015; Lyons & Beilock, 2018; see also Bulthé, De Smedt, & Op de Beeck, 2014; Eger et al., 2009; Eger, Pinel, Dehaene, & Kleinschmidt, 2015 for converging evidence using multi-voxel classification). In contrast, for number symbol processing the difference in activation patterns was not correlated with the numerical distance (Lyons et al., 2015; Lyons & Beilock, 2018; see also Bulthé et al., 2014; Bulthé, Smedt, & de Beeck, 2015 for converging evidence). The absence of distance-dependent activation patterns in these studies suggests that symbolic numbers might be represented differently — for example, as discrete, categorical units. In fact, the activation distributions for number symbols were better predicted by differences in the frequency of co-occurrence in natural speech (Lyons & Beilock, 2018). In addition, in these studies the activation distributions for the same quantity presented symbolically and nonsymbolically (e.g., the digit '6' and six dots) were not correlated, as would be expected if the same neuronal populations represented both quantity types and in the same representation format. Nonetheless, it should be noted that in these studies the total amount of activity in the same regions was still modulated by the cardinality for both number symbols and nonsymbolic stimuli. Overall, the RSA results suggest that the regions that represent symbolic and nonsymbolic numbers overlap, but the format of the representations is different and possibly neuronal populations that represent them are distinct (it is also possible that the same neuronal populations represent them, but the format of representations is different). Thus, whereas the adaptation paradigm studies supported coarse tuning in case of both number symbols and nonsymbolic quantities, results of RSA studies only support it in case of nonsymbolic quantities.

One suggested explanation for different conclusions drawn from RSA and adaptation paradigm studies is that RSA analyses are more sensitive to widely distributed activity, whereas the adaptation paradigm is more sensitive to representations at a fine spatial scale (see Eger, 2016; Wilkey & Ansari, 2019 for this suggestion). The population coding of nonsymbolic quantities possibly results in a widely distributed activity pattern which makes it relatively easily detected by RSA, whereas recognition of number symbols possibly activates on only few finely-tuned neurons which makes it less (or not at all) detectable by RSA. On the other hand, the fact that the adaptation paradigm relies on the memory phenomenon makes it more sensitive to semantic representations which are created or extracted as a result of the perceptual input processing (estimation result from an array of dots or retrieved quantity from number symbols). More research is needed to determine whether this suggestion corresponds to reality.

Together, the results reported with RSA and the adaptation paradigm allow for a possibility that number symbols represented by the same neurons as nonsymbolic quantities, but only partially recycle representations of nonsymbolic

quantities and are more efficient in encoding (see also e.g., Reynvoet & Sasanguie, 2016 for this argument). This is the possibility that we described above based on the behavioral data (section 2.2.2.3), and has previously been suggested in a computational model where the same neural network learned nonsymbolic and symbolic quantity representations, but coded number symbols with substantially sharper representations than nonsymbolic quantities (Verguts & Fias, 2004; see also Dehaene, 2007).

2.2.3.4 Interim summary

The neuroimaging research described here suggests that prefrontal and parietal areas, and specifically the intraparietal sulcus and surrounding areas, play a crucial role in both symbolic and nonsymbolic quantity processing; this follows from both brain damage patient studies and fMRI findings. There is strong evidence from single cell recordings supporting the coarse tuning hypothesis for the format of representations of nonsymbolic quantities in the brain. When it comes to number symbol representations, the evidence from different paradigms is mixed. Results based on the adaptation paradigm suggest that number symbols and nonsymbolic quantities are represented by overlapping neuronal populations and that number symbols are also represented according to the coarse tuning hypothesis. On the other hand, results of RSA studies do not support this representation format for number symbols. When it comes to the question of potentially overlapping neuronal populations representing the two formats, there is some evidence supporting this idea, but also contradictory findings, so it remains an open question.

2.3 Nonsymbolic quantities and quantifier processing

2.3.0.1 Quantifier classes

Besides number symbols, we can symbolically refer to information about quantity using natural language quantifiers such as 'some,' 'few,' 'most,' etc. Quantifiers are an integral part of human languages. Whereas number symbols refer to an exact quantity, many natural language quantifiers refer to cardinalities approximately or refer to the relations between cardinalities of (sub)sets. Different ways to decide what should be considered a quantifier have been proposed, as have different classifications of quantifiers (e.g., Barwise & Cooper, 1981; Partee, 1995; Szymanik, 2016b). For the purpose of the present review, we adopt a definition and a classification of quantifiers suggested by Keenan in a recent book reviewing quantifiers of different languages of the world (Keenan, 2012).

Broadly Keenan distinguishes *D-quantifiers* and *A-quantifiers* (following a classification based on syntactic properties, Partee, 1995), where D-quantifiers refer to those that are arguments of predicates or bind arguments of predicates ('D' stands for determiners), whereas A-quantifiers directly build predicates ('A' stands for adverbs, auxiliaries, affixes — e.g., 'always,' 'usually,' 'must,' etc.). Here, we focus specifically on D-quantifiers and omit discussion of A-quantifiers as they are less homogenous and considerably more complex when considering their relation to nonsymbolic quantities.

In the classification proposed by Keenan, *Generalized Existential (Intersective)* quantifiers refer to those quantifiers (Q) for which, given sets A and B, Q(A)(B) depends on $A \cap B$, i.e., the truth-value of the quantified expression depends on the number of As that are Bs. Within this class, *cardinal quantifiers* refer to cardinalities. Both imprecise/approximate terms such as 'some,' 'several,' 'a few,' 'a couple,' 'a dozen' and number symbols referring to exact cardinalities ('zero,' 'one,' 'two,' etc.) are included in this subclass. Here, we will focus on the imprecise/approximate generalized existential quantifiers, since those referring to exact cardinality have already been discussed as number symbols. Another subclass within generalized existential quantifiers suggested by Keenan is *value judgment quantifiers*, which refer to a given cardinality as compared to an expected cardinality — e.g., 'many,' 'few,' 'enough' (as in 'Few students attended the lecture,' where *few* refers to a quantity of students as opposed to an expected quantity). The second class, *Generalized Universal (Co-intersective)* quantifiers are those for which Q(A)(B) depends on A - B, namely the set of As that are not Bs. This class includes 'all,' 'every,' and 'each' (for all of which the set of As that are not Bs is empty). The third class, *Proportional* quantifiers refer to those for which Q(A)(B) depends on $|A \cap B|/|A|$, namely the proportion of As that are Bs. This class includes 'many,' 'few,' 'most,' 'more than half.' Note that under this classification 'many' and 'few' in some cases refer to cardinality and in

some cases refer to proportions. The final and fourth class that Keenan suggests is that of *Morpho-syntactically Complex* quantifiers, where syntactically complex quantifiers are listed. From this last class of quantifiers, we only discuss *modified numerals* such as 'more than two,' 'at least/at most five,' 'exactly/only/just ten,' etc. where an explicit number symbol is used. This type of quantifier is of special interest to us since the inclusion of number symbols necessitates that brain mechanisms for number symbol processing are involved; moreover, this type of quantifier has already been compared to other quantifiers within research into quantifier processing by the brain.

As we remarked above, the adopted quantifier classification is to a certain extent arbitrary, and alternatives exist. Quantifiers can be divided with respect to logical definability (first-order, e.g., 'all' or 'some' vs. higher-order quantifiers, e.g. 'most'), computational complexity (e.g. recognizable by finite-automata, like 'all,' or not recognizable by finite-automata, like 'an even number of'), historical reason (e.g. distinguishing Aristotelian quantifiers 'all,' 'some,' 'not all,' 'some not'), or even combinations of these various criteria (see e.g., Partee, 1995; Szymanik, 2016b). In general, it is difficult to force quantifiers into categories. Even quantifiers within the classes that we defined show substantial differences from each other. However, even though we support the idea of considering all quantifiers separately, in practice it is only possible to make progress by trying to draw some semantic generalizations. We chose Keenan's classification because it has already been used in the context of cross-linguistic research and seems to be apt for describing the human repertoire of quantifier concepts. The majority of our discussion focuses on two uncontroversial types of quantification: modified numerals and proportional quantifiers. Two other classes proposed by Keenan, generalized existential and generalized universal, need to be interpreted more carefully by asking questions about their relationship to proportional quantifiers and logical reasoning. If the research program outlined in this paper turns out to be successful then we predict that a much more fine-grained classification of quantifiers from the perspective of their relationship to symbolic and nonsymbolic quantity processing should emerge.

2.3.0.2 Relating quantifiers and nonsymbolic quantities

Let us now consider the quantifiers of each class as presented above in terms of their potential relation to nonsymbolic quantity representations and processing.

In our classification, generalized existential quantifiers like 'some,' 'several,' 'a few,' 'a couple,' 'a dozen,' etc. are considered to refer to imprecise/approximate cardinalities. The fact that they refer to imprecise cardinalities makes them compatible with nonsymbolic quantity representations in the brain — when someone refers to a quantity of objects as 'several,' we do not know what exact quantity they have in mind, just as we cannot perceive an exact quantity when presented with a set of objects and do not count them. In fact, this makes generalized

existential quantifiers more compatible with nonsymbolic quantity representations than number symbols are. This means that, unlike number symbols, these quantifiers could be direct references to nonsymbolic cardinality representations. One aspect that should be mentioned for these quantifiers is that not everyone agrees with Keenan's suggestion that generalized existential quantifiers refer to approximate cardinalities. There are suggestions that these quantifiers refer to proportion information or are at least ambiguous between the approximate cardinality and proportional readings (see e.g., Partee, 2004). This is because their meaning does not refer to any particular approximate quantity (it is not the case that, e.g., 'some' always refers to 2-5 objects), but instead, at least in some cases, they seem to be dependent on the total number of objects available in the relevant context (for supporting empirical evidence see, e.g., Pezzelle, Bernardi, & Piazza, 2018, experiment 1; Newstead & Coventry, 2000). The discussion below should be valid for both positions — if generalized existential quantifiers refer to proportions, then the discussion of proportional quantifiers applies.

Proportional quantifiers like 'many,' 'few,' 'most' are thought to refer to the ratio between two cardinalities — the cardinality of all objects in the context and the cardinality of objects that possess the relevant feature. We know that, when comparing two nonsymbolic cardinalities, behavioral performance and neuronal activation patterns are modulated by the ratio between two presented nonsymbolic cardinalities. Moreover, we know that ratio information is represented by the brain (along with cardinality information; as discussed in section 2.2.3.3) when we are presented with two nonsymbolic cardinalities. Thus, the ratio is encoded and plays a crucial role in nonsymbolic quantity processing. Proportional quantifiers could then potentially be direct references to the ratio information extracted by our nonsymbolic quantity processing system. Thus, it is possible that whereas generalized existential quantifiers refer to the extracted approximate cardinality information, proportional quantifiers refer to the extracted ratio information, both computed and made available by our nonsymbolic quantity processing system.

The link between the generalized universal quantifiers like 'all,' 'every,' 'each' and nonsymbolic quantities is less clear than in the case of other classes of quantifiers we consider here. On one view, the meaning of these quantifiers is evaluated using logical reasoning rather than the quantity system since knowledge of the quantity is not required to understand them. What is required is rather the ability to find counterexamples (e.g., if at least one object of a set does not possess a property, 'all' cannot be applied), so these quantifiers can in principle be processed independently of quantity representations (see Halberda, Taing, & Lidz, 2008; Troiani, Peelle, Clark, & Grossman, 2009; using the same argument, these researchers suggest that 'some' (which we here classify as generalized existential) does not involve quantity processing either). On the other hand, others have suggested that the ability to find at least one counterexample already entails that number processing is involved (Clark & Grossman, 2007; see also Olm, McMillan,

Spotorno, Clark, & Grossman, 2014 for a similar argument). Relatedly, different generalized universal quantifiers have different semantic functions, such as distributivity. For instance, while 'each' tends to refer to individuals and their properties, 'all' and 'every' usually refer to sets of objects. Recently acquired preliminary evidence suggests that this difference translates into variability in mental representations of the universal quantifiers (Knowlton, Pietroski, Halberda, & Lidz, 2020). Here, we do not take a position in the question whether generalized universal quantifiers should recruit nonsymbolic quantities, but only highlight it. Multiple studies discussed below included an investigation of processing of specifically the quantifier 'all'.

Finally, modified numerals such as 'more than two,' 'at least/at most five,' etc. are relevant for the present discussion because they include number symbols. These quantifiers require that a person has learned to operate with exact number symbols. When considering the involvement of brain mechanisms, those processing number symbols have to get involved in order for these quantifiers to be understood and produced. Since we know relatively a lot about number symbol processing, we have specific predictions about the mechanisms that should be involved in their processing (for example, at the neuronal level we expect to observe the involvement of neuronal populations in the intraparietal cortex). In this sense, this class of quantifiers will sometimes function as a good baseline for seeing the involvement of quantity processing mechanisms in the case of other quantifiers.

2.3.0.3 Differences between quantifiers and number symbols in relation to nonsymbolic quantity processing

We know that some languages have an upper limit to number words that exist to refer to exact cardinalities: some languages have number words only up to 3-5, some have a number higher than 5 as an upper limit, and a few are even reported to have an upper limit of 'one' or 'two' (e.g., Bowern & Zentz, 2012; Epps, Bowern, Hansen, Hill, & Zentz, 2012; see Carey & Barner, 2019; Núñez, 2017 for review and references). Thus, the symbolic number system (at least to the extent that Western cultures use it) does not arise during the course of human life spontaneously, but rather requires explicit training. In contrast, regardless of the extent of the numerical system of a language, all languages seem to have words to refer to approximate cardinalities by means of quantifiers, analogous to e.g., 'some,' 'several,' 'few,' 'many' in English (Bowern & Zentz, 2012). We also know that understanding and communicating using quantifiers does not require explicit training because children are able to use them before they start math education (e.g., Barner, Chow, & Yang, 2009; Barner, Libenson, Cheung, & Takasaki, 2009; Dolscheid et al., 2017). Finally, in cultures that do have an extensive number symbol system, quantifiers are still used in communication even if the exact number of objects is known (e.g., someone saying that they 'bought several books'

even though they know that they bought exactly three books). Given these considerations, quantifiers can be seen as a more natural way to refer to nonsymbolic quantity information in human languages than number symbols (see also e.g., Clark & Grossman, 2007; Coventry, Cangelosi, Newstead, & Bugmann, 2010, for this suggestion). Consistent with the possibility that quantifier processing is based on the nonsymbolic quantity processing mechanisms outlined above, speakers of all languages perform equally well when it comes to nonsymbolic quantity perception and comparison (Ferrigno et al., 2017; Gibson, Jara-Ettinger, Levy, & Piantadosi, 2017; Pica et al., 2004).[10]

Another substantial difference between quantifiers and number symbols is the context-sensitivity of quantifiers (which is additional to the imprecise nature of the quantity to which they refer; see also Moxey & Sanford, 1993; Newstead & Coventry, 2000 for this point). While the number symbol 'two' always refers to a cardinality 'two,' there is no fixed cardinality or proportion for quantifiers. Possible exceptions to this are generalized universal quantifiers ('each,' 'every,' 'all'), but even here the exact quantity that 'all' means is in a sense different (i.e. 'all' refers to a different quantity for a group of 5 objects than for a group of 10 objects). Rather, the quantity that these quantifiers refer to depends on a typical quantity for an object that they refer to (e.g., 'many,' when referring to 'pandas' compared to 'ants,' will mean a different quantity), on the expected quantity for a particular situation, on specific speaker experiences (Heim et al., 2015; Ramotowska, Steinert-Threlkeld, Leendert, & Szymanik, 2020; Yildirim, Degen, Tanenhaus, & Jaeger, 2016) and possibly other factors.[11] The context-sensitivity of quantifiers makes them, again, more compatible with nonsymbolic quantity representations than number symbols are. We know, for example, that there are individual differences in performance with more difficult ratios in nonsymbolic quantity comparison tasks (what is typically referred to as *nonsymbolic number acuity*; e.g., Halberda & Feigenson, 2008), that there are individual differences in underestimation bias in estimation tasks (as discussed in section 2.2.2.1; Crollen

[10] A related point is that while it is, in principle, possible to have a one-to-one correspondence in terms of quantity between quantities represented nonsymbolically and by number symbols (e.g., six dots and the digit '6'), it is not possible to find clear correspondence between quantifiers and nonsymbolic quantities in this way. It might therefore seem like quantifier meanings are less comparable to nonsymbolic quantities than number symbol meanings are. However, considering that nonsymbolic quantities beyond the subitizing range are not exactly represented anyway, this correspondence is not useful from the perspective of questions about shared or distinct brain processing and representations. Again, in this sense quantifiers seem to be a more natural reference to nonsymbolic quantity information than number symbols are.

[11] For proportional quantifiers, part of context-sensitivity can be potentially explained by the fact that they refer to a proportion that is invariant to absolute quantities. However, this still does not explain context-sensitivity in terms of speaker differences — different people have different internal criteria for what proportion should be considered 'many ants' (perhaps for a person with an insect phobia just three ants would be sufficient; moreover, we know that the people's internal thresholds can also change in a course of a conversation (Heim et al., 2015; Ramotowska et al., 2020; Yildirim et al., 2016).

et al., 2011), and that estimates of cardinality of object arrays are influenced by how elements are clustered together and spatially organized within a visual scene (Im, Zhong, & Halberda, 2016). As far as we know, the connection between the context-sensitivity properties of quantifiers and the related specific features of nonsymbolic quantities has not been yet studied in the literature. Here, we do not attempt to relate the context-sensitivity properties of quantifiers to specific features of nonsymbolic quantity processing, as that would make the present effort unmanageable, but only note these properties and leave them for future research.

Related to context-sensitivity is the need to *choose* an appropriate quantifier to describe a certain quantity. This involves not only deciding whether, e.g., the given proportion should be considered low, but also which of a variety of similar-in-meaning quantifiers should be used (e.g., 'few,' 'several,' or 'some'). This means that decision-making processes will be involved in producing a quantifier — unlike in the case of number symbols, where there is only one corresponding symbol.

Finally, in contrast to number symbols, different quantifiers will lead to different inference patterns when interpreting them — e.g., if 'some people ate oranges' is true, then 'some people ate' has to be true as well.[12] Downward monotone and upward monotone quantifiers are traditionally distinguished (Barwise & Cooper, 1981; this property is also referred to as *quantifier polarity*). This aspect is traditionally seen as purely linguistic (i.e., not involving quantity processing systems). While decision-making and inference licensing properties of quantifiers are important, in this review we do not try to fully cover them; they require a thorough consideration on their own. We consider these linguistic and decision-making processes as always additional to the quantity processing that takes place for quantifiers.

In the rest of this section, we consider the existing evidence and suggest future research questions for whether and how cognitive systems supporting nonsymbolic quantity are involved in processing natural language quantifiers. The main question here is whether and to what extent the same representations and processing mechanisms are involved in quantifier and nonsymbolic quantity processing by the brain. We review all major studies to date investigating this relation for quantifiers of which we are aware. Where possible, we draw parallels with evidence from research into number symbol processing. Importantly, quantifiers might be linked to nonsymbolic quantities to a larger extent than number symbols are, since they have a set of different properties, sometimes better aligned with the properties of nonsymbolic quantities (as discussed above).

[12]In contrast, if 'five people ate oranges' is true, it is not the case that then 'five people ate' has to be true (there could be additional people present in the context who ate things other than oranges). There is a debate (see e.g., Hurewitz, Papafragou, Gleitman, & Gelman, 2006, for a short review) about whether the number symbols in this context are interpreted as 'at least five (people ate oranges),' in which case the number symbol meaning would be upward monotone. We will leave this debate aside, however.

2.3.1 The developmental perspective

We know that even pre-linguistic infants are able to distinguish nonsymbolically presented quantities and that their ability to discriminate improves with development, allowing increasingly smaller ratios to be distinguished (e.g., Izard et al., 2008; Spelke, 2011; Wynn, 1998; Xu & Spelke, 2000). Children are also able to understand some quantifiers from approximately the age of two (e.g., Barner, Chow, & Yang, 2009; Barner, Libenson, et al., 2009). Parallel to the hypotheses about number symbol learning, one hypothesis about quantifier learning would be that since children have the nonsymbolic quantity processing system available, they simply associate or map the quantifier meanings onto these nonsymbolic quantity representations. An alternative hypothesis is that quantifier comprehension and production develop as a separate system, not relying on nonsymbolic quantities. In the case of specifically generalized universal quantifiers, recall that logical reasoning might be especially important, so children would need to develop this first in order to correctly understand and use these quantifiers. For example, children understand the meaning of 'some' at an adult level (i.e., interpret it as adults do) at a later point in development than they understand 'all' (Barner, Chow, & Yang, 2009; Barner, Libenson, et al., 2009; Dolscheid et al., 2017). We will not discuss the development of logical reasoning in detail here, keeping our focus solely on the question of the relation to the nonsymbolic quantity processing system.

2.3.1.1 Learning quantifiers by mapping them to nonsymbolic quantities

To look at the interface between quantifiers and nonsymbolic quantities in young children, Odic and colleagues (Odic, Pietroski, Hunter, Lidz, & Halberda, 2013) used a sentence-picture verification task with the comparative quantifier 'more'. Eighty children aged two to four years were asked to verify the statement 'are more of these dots blue or yellow?' (as well as the statement 'is more of the goo blue or yellow?' in another condition for which they observed the same result). They reasoned that if children use their nonsymbolic quantity processing system to evaluate whether the quantifier fits as the description, they should observe the typical psychophysical pattern of ratio-based performance for nonsymbolic quantity comparison seen in adults, albeit given more noisy representations. Children performed above chance in this task at approximately age 3.3. Those children who succeeded indeed showed a pattern of performance consistent with nonsymbolic quantity processing. Odic and colleagues interpret their results as suggesting that 'more' interfaces with perceptual quantity processing mechanisms, and that children have access to this interface as soon as they understand the meaning of the comparative 'more'.

If quantifiers indeed rely on nonsymbolic quantity processing, one could ex-

pect children who perform nonsymbolic quantity comparison at a higher level (i.e., who are able to distinguish smaller ratios) to also have a better understanding of quantifiers. The only study to date we are aware of that investigates this question looked at the correlation between the two abilities. Dolscheid and colleagues (Dolscheid et al., 2017) asked 39 children aged between three and six years old to give a number of objects corresponding to one of eight German quantifiers ('Can you put all/a/none/both/most/many/some of the bananas into the bowl?'). The children's performance in this task (assessed based on whether they gave a quantity in the range matching that of adult control participants) was overall correlated with the ratio they were able to discriminate in a nonsymbolic comparison task. This correlation was significant when controlling for age, IQ, and the children's level of knowledge of number symbols. However, when investigated more closely based on performance with individual quantifiers, only the quantifiers 'both' and 'most' were related to performance on the nonsymbolic quantity comparison task. The fact that only two quantifiers were clearly related to nonsymbolic quantity performance is unexpected given that among those given to the children, at least quantifiers 'many' and 'some' can be thought of as those that should be related to nonsymbolic quantities. A potential explanation may lie in the fact that the average age of children who participated in this task was 4.5 years old. These children have likely already mastered other quantifiers rather well (surpassing the initial reliance on purely nonsymbolic quantities) and perhaps showed ceiling performance that did not allow for correlations to arise. Indeed, when examining the performance for each quantifier it becomes apparent that they perform at ceiling for all quantifiers except for 'most,' 'both,' and 'some'.[13] This explains why for other quantifiers there was no relationship, though it still does not answer why there was no relationship with 'some.'

2.3.1.2 Order of acquisition of quantifiers

As discussed in section 2.2.1.2, we know that the meaning of number symbols is acquired in a particular order — 'one' through 'four' sequentially, followed by an understanding of the cardinality principle. A parallel question for quantifiers would be whether there is any particular universal order of acquisition of quantifiers by children learning different languages. Katsos and colleagues (Katsos et al., 2016) suggest that if quantifiers, like number symbols, are acquired in order of increasing cardinality, it follows that 'a few' and 'some' should be acquired earlier in development, whereas 'most' and 'all' should be acquired later in development. This prediction is not borne out given the observation that children as early as two years old understand 'all,' but even some 7-year old children have not yet fully acquired the meaning of 'most' (see e.g., Barner, Chow, & Yang,

[13]Interestingly, children learning English show a parallel pattern, with 'both' and 'most' being the most difficult quantifiers to acquire (Barner, Chow, & Yang, 2009); see also (Sullivan, Bale, & Barner, 2018) for evidence that 'most' might not be fully acquired until later in childhood.

2009). Instead, given that quantifiers are richer in meaning (due to the inference patterns they give rise to), Katsos and colleagues suggest that there are constraints in quantifier learning that are absent in learning number symbols. They present four such constraints (given the monotonicity, totality, complexity, and informativeness properties of quantifiers) based on which they make predictions for quantifiers corresponding to the English 'all,' 'none,' 'some,' 'some...not,' and 'most.' Katsos and colleagues collected data from children learning 31 different languages (all languages were those of industrialized societies with complete number symbol systems). Children learning most of these languages conformed in their performance to predictions based on each of their proposed constraints. Katsos and colleagues, therefore, suggest that the order of acquisition of quantifiers is driven by properties that can be characterized as something like 'semantic complexity' rather than the cardinalities to which they refer.

2.3.1.3 Questions from the developmental perspective

One set of questions regarding the acquisition of quantifiers concerns the (availability of the) interface between quantifier comprehension and perceptual systems of nonsymbolic quantities in sentence-picture verification. The only such study with children was conducted by Odic and colleagues (Odic et al., 2013). This study suggests that children make use of nonsymbolic quantity representations to evaluate 'more' as soon as they understand the comparative meaning of 'more'. This observation needs to be confirmed in replications. In addition, follow-up research should investigate whether this generalizes to other quantifiers such as 'some,' 'several,' 'many' etc. If it does, what kind of information do children then extract from nonsymbolic quantity representations using this interface for each of the quantifiers and do they change over the course of development? A relevant fundamental question about whether this paradigm really taps into quantifier knowledge is discussed below; see section 2.3.2.5 where we discuss evidence obtained this paradigm with adults.

Katsos and colleagues (Katsos et al., 2016) argue that cardinality does not play a role in the order of acquisition because children do not master quantifiers in the order of the cardinality or proportion to which they refer. However, there is an alternative hypothesis about the order of the acquisition of quantifiers that can be derived from what we know about the development of nonsymbolic quantity processing. We know that children improve in their ability to distinguish between two nonsymbolic quantities in the course of development — their estimates become more accurate and they learn to distinguish increasingly smaller ratios (e.g., Feigenson, 2007; Halberda & Feigenson, 2008). Perhaps predictions about the order of acquisition should be related to how well children can distinguish between pairs of quantifiers at a given developmental stage rather than to the specific cardinalities to which each quantifier refers. The further apart two cardinalities or proportions are from one another, the sooner children would

be able to successfully distinguish them perceptually and, therefore, the earlier they will master the difference between the corresponding quantifier pairs. This proposal predicts, for example, that children will successfully distinguish between 'few' and 'many' at an earlier point in development than they successfully distinguish between 'few' and 'several' or between 'many' and 'most'. While Katsos and colleagues present convincing evidence that semantic complexity plays a role in the order of acquisition, it is possible that the development of nonsymbolic quantity representations plays a role in the order of acquisition alongside these factors. Note also that whether order-of-acquisition accounts are able to predict the order of acquisition of all' depends on whether we consider generalized universal quantifiers to also rely on the nonsymbolic quantity system, leaving it a question for debate.

Only one study to date has examined whether there is a potential correlation between nonsymbolic quantity discrimination performance and quantifier knowledge (Dolscheid et al., 2017), in one language and with a sample of 39 children. Studies with a larger sample and age range of children as well as with different languages are needed to see if this relationship exists. Moreover, as we have observed, it would also be important to break down the relationships by specific quantifiers or quantifier classes. In addition, in analogy to studies on the relationship between number symbols and nonsymbolic quantities, one could also look into whether training participants to discriminate nonsymbolic quantities improves their performance with quantifiers. While such training results in little or no improvement of performance in number symbol tasks (as discussed in section 2.2.1.1), if natural language quantifiers rely on nonsymbolic quantity representations, this may result in improved performance with quantifiers.

2.3.2 The behavioral perspective

In estimation tasks, we have seen that people are able to give an approximation of cardinality using number symbols. The fact that number symbols refer to exact cardinalities makes it possible to find one corresponding number symbol for any particular nonsymbolic quantity after counting. In contrast, due to their imprecise meaning, there are no unambiguous, objective nonsymbolic counterparts for generalized existential and proportional quantifiers. The first question asked from the behavioral perspective is, thus, which criteria people use to decide whether a quantifier is a good description of a certain cardinality. These studies ask whether a particular quantifier corresponds to a particular cardinality or ratio in a nonsymbolic quantity representation. In parallel to estimation tasks with number symbols, where participants were asked to give number symbols corresponding to the cardinality of an array of objects, here participants were asked to give a quantifier to describe the cardinality.

When evaluating the fit between the meaning of a quantifier (at least in the case of generalized existential and proportional quantifiers) and a particular visual

scene consisting of an array of objects, clearly at least two processes have to take place — retrieval of the meaning of the quantifier and assessment of the cardinality using the nonsymbolic quantity processing system. A number of studies have looked into what kind of information about the nonsymbolic quantity is extracted and assessed in relation to the proportional quantifier meaning. These studies aim to characterize *the interface* between quantifier meaning and nonsymbolic quantity representations.

For number symbols, one prominent research direction has been investigating whether the format of number symbol representations is similar to that of nonsymbolic quantity representations. Again, this question can also be asked for quantifiers, and we will review the existing evidence to date below. In parallel with the question for number symbols, one can ask whether quantifiers are represented in nonsymbolic quantity-like, noisy and overlapping representations format or as discrete entities. Note that what complicates the picture for quantifiers is that, unlike in the case of number symbols, there is no strict linear order for all quantifiers in terms of the cardinality or proportion to which they refer. Several studies look into the underlying dimensions behind quantifiers, and we briefly touch upon these.

Because quantifiers have additional pragmatic/linguistic features in comparison to number symbols, we can ask whether these properties can influence nonsymbolic quantity processing. Assuming that there exists an interface between quantifiers and the nonsymbolic quantity representation system, one possibility is that certain quantifiers influence the comparison process in the quantity processing system when extracting information. We discuss one specific proposal for such top-down influence below.

Before we discuss behavioral (and neuronal-level below) research with quantifiers, it should be noted that whereas research on number words was mostly focused on representations of number symbols (i.e., what is stored in our cognitive system for each number symbol), in the case of quantifiers questions have also been asked regarding the processes involved in interpreting a particular quantifier, i.e., about their dynamic evaluation by various mechanisms of the cognitive system.

2.3.2.1 Mapping between quantifiers and nonsymbolic quantities

Parallel to estimation tasks with number symbols, one can also look at estimation tasks with quantifiers where people are presented with a visual array of objects and asked to produce or choose a quantifier that best describes a target set of objects (e.g., red dots or red dots surrounded by dots of a different color). Such tasks have been used to determine a cardinality or a proportion to which each quantifier refers, but for the most part they have only revealed the enormous context-sensitivity of generalized existential and proportional quantifiers with respect to the nouns with which they combine, the situational context, and indi-

vidual speaker judgments (e.g., Coventry et al., 2010; Heim et al., 2015; Moxey & Sanford, 1993; Newstead & Coventry, 2000; Yildirim et al., 2016). These aspects make it difficult to pinpoint any particular reference in terms of proportions or approximate cardinalities for each quantifier.

A recent such study from Pezzelle and colleagues (Pezzelle et al., 2018, Experiment 1) presented participants with visual displays of two types of objects (e.g., five hedgehogs and 15 balls in one scene) and asked them to choose an appropriate quantifier from a range of alternatives to describe one of the sets. Pezzelle and colleagues wanted to determine the factors that influence which quantifier is picked as the best description. To do so, they ran a regression analysis with a number of potentially relevant variables (cardinality of targets, cardinality of non-targets, subitizing/nonsubitizing range, average size of targets, average size of non-targets) for each of the quantifiers they tested. Specifically, they tested some proportional ('most,' 'many'), some generalized existential ('some,' 'few,' 'none,' 'almost none') as well as a generalized universal ('all') quantifiers. For all quantifiers they tested, except for 'almost none,' the proportion of the target items in the set of all items was the best predictor of the choice of the quantifier as the appropriate description.[14] Therefore, all these quantifiers seemed to have been interpreted as proportional in this experiment. This result could be ascribed to the nature of the task — participants always saw displays of two sets of objects (which could have encouraged their comparison) and proportional quantifiers were intermixed with others (which could bias them to viewing all quantifiers as proportional). Nonetheless, the results are interesting in terms of the classification we adopt in this paper. Specifically, they support the view that generalized existential quantifiers can refer to proportions at least in some contexts. In addition, it is surprising that the generalized universal 'all' was dependent on the proportion; this supports the possibility that quantity processing plays a role in this class of quantifiers at least in some contexts.

The study by Pezzele and colleagues (Pezzelle et al., 2018) also analyzed the particular proportions that participants associated with each quantifier. There was substantial overlap between the proportions to which quantifiers referred (e.g., when the target objects constituted 20% of all objects on the screen, participants chose 'few,' 'the smaller part,' or 'almost none' to describe their cardinality). It was nonetheless possible to order quantifiers in terms of their preferred proportion ranges or most preferred proportion. The resulting order was: 'none,' 'almost none,' 'few,' 'the smaller part,' 'some,' 'many,' 'most,' 'almost all,' 'all'. Interestingly, the range of preferred proportions was smaller and there was less overlap for low-magnitude quantifiers (i.e., quantifiers referring to smaller pro-

[14]Note that when the analysis was restricted to trials with target object quantities within the subitizing range, for 'few,' 'none,' and 'almost none' the best predictor was the number of target objects. We do not discuss this analysis since the subitizing range was defined based on the number of target objects alone, meaning that in fact the total number of objects on the display was outside the subitizing range.

portions) than for high-magnitude quantifiers. This means that low-magnitude quantifiers had relatively more specific meanings. We return to this point later (section 2.3.2.3).

2.3.2.2 The interface between quantifiers and nonsymbolic quantities

A number of studies investigate whether and how quantifiers recruit or interact with nonsymbolic quantity representations and processing mechanisms in sentence-picture verification tasks. In these studies, participants are required to understand the meaning of a sentence with a quantifier (e.g., 'Many of the dots are blue') and, subsequently, decide whether the presented visual display matches the description. Therefore, it is assumed that participants in this task process the visual display with the particular goal of extracting specific information required by the particular given quantifier meaning.

In two studies with the proportional quantifier 'most,' participants were asked to answer the question 'Are most of the dots yellow?' (or 'blue' in the second study; Lidz et al., 2011; Pietroski et al., 2009). Participants saw visual displays with dots of two or more colors for 150-200 ms. Given the restriction in the time for which the visual arrays were displayed, participants were prevented from counting. In each trial, they answered the same question by pressing 'yes' or 'no.' Within the visual displays, the ratio of dots of the target color and non-target colors was varied; presented ratios were 1:2; 2:3, 3:4, 4:5, . . , 9:10. In these studies, the accuracy of the participants' responses varied according to the ratio, mirroring the performance that would be expected in the case it was simply a nonsymbolic cardinality comparison task. Such results suggest that ratio information from the nonsymbolic quantity representation system is indeed extracted in order to evaluate fit against the meaning of the quantifier 'most'. In addition, the authors of these studies interpret the results as showing that the canonical meaning of 'most' is inherently rooted in the nonsymbolic quantity representation system. One point of criticism of these studies is that participants saw 350-400 trials with the aim of verifying exactly the same sentence. It is thus not necessarily the case that participants were retrieving the meaning of 'most' with every trial. Participants might as well just have been instructed to compare the cardinality of the dots of two different colors in a purely perceptual experiment. The second important point is that participants in these studies had only 150 or 200 ms to view the visual displays. Ideally, we would like to know whether the quantity-processing system is involved for longer viewing times as well or whether it was simply an artefact of this particular set-up.

Another sentence-picture verification study with quantifiers by Deschamps and colleagues (Deschamps, Agmon, Loewenstein, & Grodzinsky, 2015; see also Heim et al., 2012 for a similar set-up and results) avoided the issues of using a single sentence across the whole experiment as well as a short duration of visual display presentation. In this study, participants were presented with the quan-

tifiers 'more/less than half,' 'many,' 'few,' 'more/less [...] than [...]' in sentences (e.g., 'Many of the circles are yellow'). Each trial started with an auditory presentation of the sentence to be judged, and followed with a visual display. The visual displays contained circles of two colors, with the ratio between the cardinalities of the circles of each color being manipulated. The visual displays stayed on the screen for 1100 ms. The performance of participants in terms of accuracy as well as reaction time was modulated by ratio for all quantifiers. Thus, here the authors obtain the same effects while the instructions differed in each trial and the duration of the visual display was longer.[15]

Yet another similar study, by Shikhare and colleagues (Shikhare, Heim, Klein, Huber, & Willmes, 2015), also manipulates the sentence with the quantifier that is to be evaluated between trials and presented visual stimuli for 1000 ms. Shikhare and colleagues asked participants to verify sentences with the proportional quantifiers 'many' and 'few' (e.g., 'many/few of the circles are yellow') and modified numerals 'at least n' and 'at most n' (e.g., 'at least/at most seven of the circles are yellow'). For the proportional quantifiers, they also observed slower responses and more errors with smaller ratios. Of special interest are the conditions with the modified numerals since here participants compared number symbols and nonsymbolic quantities, but with an additional direction of comparison/instruction given by the 'at least/at most' quantifier. When the actual quantity of the dots of the corresponding color on display was closer to the reference number (e.g., 8 as opposed to 12 circles displayed for the sentence 'at least seven of the circles are yellow'), the reaction times were longer and accuracy was lower. Therefore, the ratio effect was also preserved here. We discuss further findings in these conditions in section 2.3.2.4 below, in relation to whether quantifiers bias the lower-level quantity comparison process.

Finally, another series of sentence-picture verification studies compares processing times for proportional and other quantifier classes. In these studies, the duration of the visual display was long enough to allow counting if participants so wished. Comparing reaction times, the studies find that participants are fastest for the quantifiers 'all' and 'some,' followed by modified numerals ('less than eight,' 'more than seven') and, finally, with the proportional quantifiers 'more than half' and 'less than half' taking the most time (Szymanik & Zajenkowski, 2010). Moreover, schizophrenic patients fell behind control subjects, in terms of accuracy, only on proportional quantifiers (Zajenkowski, Stył a, & Szymanik, 2011). Furthermore, the numerical distance between the two cardinalities to be compared in the case of proportional quantifiers influences verification time and accuracy (Zajenkowski, Szymanik, & Garraffa, 2014). Szymanik and colleagues suggest that proportional quantifiers take longer to evaluate because they involve

[15]While the authors do not explicitly mention that the order of presentation of trials with different quantifiers was randomized, this is implicit in the arguments that they make in the paper. We thus infer that the trials with different quantifiers were intermixed.

comparisons of the cardinalities of two sets, requiring the involvement of working memory and executive processes, whereas 'all' and 'some' do not require such comparison (see Szymanik, 2016a for an overview).

2.3.2.3 The representation format of quantifiers

It has been suggested that the representation format for number symbols in the human brain parallels that for nonsymbolic quantities. We can also base a hypothesis about the quantifier representation format on nonsymbolic quantity representations. Specifically, quantifiers may be organized in a network of ordered, noisy, overlapping units where those referring to larger approximate cardinalities or proportions (high-magnitude quantifiers) have more overlap with each other than those referring to smaller approximate cardinalities or proportions (low-magnitude quantifiers). Recall that evidence for such a format has been reported not only for approximate cardinalities, but also for the ratio information (Jacob & Nieder, 2009a; Jacob et al., 2012; see section 2.2.3.3 above). Therefore, such a representation format is at least possible for generalized existential (which we think refer to cardinalities or proportions) and proportional quantifiers (which we think refer to proportions). The representation format of generalized universal quantifiers is more tricky because, as discussed, it is not clear to what extent they are related to quantity processing rather than logical reasoning. On the other hand, possibly even these quantifiers rely on quantity information and there is some empirical evidence to suggest that they are also understood as referring to proportions (as discussed above in relation to the empirical results observed by Pezzelle et al., 2018). Alternatively to the representation format mirroring nonsymbolic quantity, quantifiers may be organized as discrete entities, not in linear order and without any overlap in meaning representations due to quantity or cardinality reference overlap. In such a network, each quantifier representation would be separate from others, not competing for activation due to overlap (but possibly still competing for activation for other reasons).

Importantly, unlike for number symbols, quantifiers have prominent features in addition to their reference to quantity — they are context-sensitive, give rise to different pragmatic inference patterns, and some are in a special antonym relation to each other (e.g., 'many' vs 'few'). Their representations, thus, should contain more information than simply reference to quantity and they might be organized along more than one dimension (Pezzelle et al., 2018; Routh, 1994), forming multiple different networks.

To investigate the features that comprise quantifier representations in the human cognitive system, another experiment in the above-mentioned study by Pezzelle and colleagues (Pezzelle et al., 2018, Experiment 2) asked participants to evaluate the semantic similarity of pairs of quantifiers on a scale from 1 to 7. They then used multidimensional scaling to look for underlying dimensions that would explain the judgments of similarity. The results indicated that just two

dimensions presented a rather good fit for their data (R^2=.988), where one dimension seemed to correspond to a separation between the low- and high-magnitude quantifiers they used ('none,' 'almost none,' 'few,' 'the smaller part' vs. 'many,' 'most,' 'almost all,' 'all') and the second dimension distinguished between the low-magnitude quantifiers themselves while not distinguishing between the high-magnitude quantifiers. This suggests that the overlap between representations of low-magnitude quantifiers is substantially lower than the overlap between representations of high-magnitude quantifiers, as would be expected from an organization format similar to that of nonsymbolic quantity representations. Taken together with the results of their other experiment in the same study (discussed above), which showed less overlap in distributions of proportions that were judged to correspond to lower-magnitude quantifiers, the data from Pezzelle and colleagues supports the hypothesis about the existence of ordered representations of quantifiers with more overlap for quantifiers denoting larger proportions.

2.3.2.4 Quantifiers biasing the nonsymbolic quantity processing mechanism

Another question asked about quantifiers concerns the potential top-down influence of specific quantifiers on nonsymbolic quantity perception or comparison processes. Specifically, do we perhaps employ different mechanisms/strategies for quantity comparison when quantity information is extracted by different quantifiers? Shikhare and colleagues (Shikhare et al., 2015) suggest that quantifier semantics does indeed bias quantity processing mechanisms. Let us take the example they give of comparing an array of 5 dots against a modified numeral — 'at least seven' where the key will be 'at least.' They argue that in order to perform this comparison, we need to activate a quantity distribution corresponding to the reference quantity 'seven' and compare it to the observed quantity 5. However, because 'at least' typically focuses our attention on larger quantities than the reference (e.g., 'at least seven' is typically used to mean 'seven or more'), the quantity distribution of 'seven' will be skewed towards larger quantities; if we imagine the quantity representations in a left-to-right direction, it will have a right skew. We are therefore comparing a uniform distribution around 5 to a right skewed distribution around 7. Because the distribution for 7 is right skewed, there will be less overlap with the distribution for 5 than if both distributions were uniform. This should result in faster reaction times and higher accuracy for 'at least seven' than if we were to simply compare the quantities 5 and 7. Thus, the ratio effect will be different to that of a case where two quantities are compared without the quantifier biasing the comparison process. The opposite should be the case for 'at most seven,' since 'at most' focuses our attention on smaller quantities than the reference (e.g., 'at most seven' is typically used to mean 'a maximum of seven, or less').

Another example of the potential influence of quantifier semantics on the

comparison process has to do with the contrast between 'most' and 'more than half'. It has been suggested in the literature and supported by experiments that the two quantifiers have roughly the same extension, i.e., both mean 'more than half.' Hackl (2009) has observed that these quantifiers are potentially associated with different information extracted from the quantity processing system: while 'more than half As are B' involves dividing the total number of A's in half, verifying 'most As are B' requires comparing the total number of A's that are B's with the number of A's that are not B's. However, others have failed to replicate these results and instead suggest that the different roles that the working memory plays in the verification of each of these quantifiers as well individual differences in the use of various cognitive strategies are a better explanation for the difference that Hackl observes (Steinert-Threlkeld, Munneke, & Szymanik, 2015; Talmina, Kochari, & Szymanik, 2017). Independently, Solt (2016), using corpus data, suggests that whereas 'most' can be used when only approximate cardinality information is available, 'more than half' can only refer to the result of a precise comparison, so it possibly relies on symbolic number processing. Recently, Ramotowska and colleagues (Ramotowska et al., 2020) have applied new modeling techniques to the verification data and discovered that mental representations of 'most' (operationalized as thresholds separating true and false instances) vary across subjects and affect the verification process. However, these effects are not present for 'more than half'. Summing up, these debates leave us with two possibilities: either 'most' and 'more than half' have the same meaning but interact differently with the nonsymbolic quantity system or they subtly differ in meaning.

Consistent with particular quantifiers biasing the quantity comparison mechanisms, Deschamps and colleagues (Deschamps et al., 2015) found a difference in performance between evaluating a phrase with a quantifier as opposed to the same meaning being conveyed using a mathematical symbol (for example, 'Many of the circles are blue' as opposed to instructions given as a depiction of a blue square followed by a sign ' $>$' and followed by a yellow square, the alternative color of circles in the visual display). Whereas error rates and reaction times were different for pairs of antonymous quantifiers despite the only difference being in the direction that they referred to (e.g., 'many' vs 'few,' possibly due to bias introduced by each quantifier, as suggested by Shikhare and colleagues; see Deschamps et al., 2015 for an extensive discussion of other possible explanations), no such difference was observed for the two opposite mathematical symbols. The fact that simply giving instructions using a quantifier resulted in a different performance speaks to the idea that the quantifier did somehow influence or bias the comparison process.

2.3.2.5 Questions from the behavioral perspective

The reviewed studies that looked at the interface between quantifiers and nonsymbolic quantity system (Deschamps et al., 2015; Heim et al., 2012; Lidz et al., 2011; Pietroski et al., 2009; Shikhare et al., 2015) as well as the earlier described study by Odic and colleagues with children (Odic et al., 2013) all report the same ratio-dependent performance even though they investigate different proportional quantifiers — 'most,' 'more/less than,' 'many,' 'few,' 'more'. The interesting question here is whether by interfacing with quantity processing mechanisms these quantifiers simply extract the ratio between two sets, which is then subsequently used to make a decision regarding whether the quantifier is applicable. If this is the case, as the evidence seems to say, any differences between their meanings should not be due to quantity processing but to specific extracted ratio values (e.g., for the difference between 'many' and 'few') and possibly some other properties (e.g., inference patterns, pragmatic aspects etc.; e.g., in the case of 'many' and 'most,' where we intuitively believe there is a difference). Alternatively, however, it could be said that the set-up of these tasks was such that participants did not in fact evaluate quantifier meanings but performed a perceptual judgment — simply chose the larger/smaller quantity set. In this case a better task would be required to allow us to observe the differences between the kind of information that is extracted from quantity representation mechanisms. One possibility, for example, instead of an experiment with many trials where participants may develop and adjust strategies, is to administer few items but with many participants (this has been done, e.g., by Register, Mollica, & Piantadosi, 2020). Another possibility would be to compare performance in the set-up with a visual scene and a set-up without a visual scene (e.g., as has been done by Schlotterbeck, Ramotowska, van Maanen, & Szymanik, 2020).

Furthermore, in terms of the interface between quantifiers and nonsymbolic quantity, it remains to be seen whether generalized existential quantifiers such as 'some,' 'several,' 'a few,' 'enough,' 'a couple,' 'a dozen,' etc. as well as generalized universal quantifiers such as 'all,' 'every,' 'each' also interface with quantity processing systems, and if they do, what kind of information they extract. For the generalized existential quantifiers we would expect cardinality information to be extracted, whereas for the generalized universal quantifiers it is more difficult to make predictions. In fact, researchers are starting to look at the interface between generalized universal quantifiers and nonsymbolic quantities. In a manuscript under review at the time of writing, Knowlton and colleagues (Knowlton et al., 2020) present experimental results suggesting that verifying sentences with the quantifiers 'all' and 'every' against visual displays triggers a representation of the cardinality of a set, whereas 'each' does not.

We suggested two possibilities for the representation format of quantifiers — linearly ordered overlapping representations with increasing overlap along with increasing quantities or proportions (parallel to nonsymbolic quantities) or a

network of discrete items. The evidence presented by Pezzelle and colleagues (Pezzelle et al., 2018) supports the former, but these results are based on explicitly requested similarity judgments for which participants used their intuition. To gather further evidence, some of the paradigms used to investigate the number symbol representation format can also be used with quantifiers. Specifically, matching and priming tasks could be used, where different pairs of quantifiers instead of pairs of number symbols would be presented. If their representations overlapped, quantifiers more similar to each other in meaning would be more difficult to distinguish, resulting in longer RTs and lower accuracy, and would prime each other more. As mentioned, however, quantifier representations potentially contain features other than quantity information. These aspects should be taken into account in designing experiments and interpreting results.

A related question is why quantifier meanings/representations should overlap more for increasing quantities or proportions at all. We have a suggestion for this that could be explored in future work. Low-magnitude quantifiers refer to larger ratios between target objects and the total number of relevant objects (e.g., 'few' referring to 3 out of 10 items, ratio 3:10), whereas high-magnitude quantifiers refer to smaller ratios between target objects and the total number of relevant objects (e.g., 'many' referring to 7 out of 10 items, ratio 7:10). Since our nonsymbolic quantity-representation system is more accurate with larger proportions, it is also more capable of supporting quantifiers referring to larger proportions. Larger ratios overlap less and remain sharp, so they result in less confusion and fewer errors. On the other hand, the meaning of quantifiers referring to smaller ratios is blurry/imprecise because our nonsymbolic quantity representation system is not capable of perceiving these differences to the same extent. To look at this question, a learning computational model could be used that would start with equal overlap for lexical items referring to cardinalities and ratios across the whole range, and with a system where quantity representations have properties of nonsymbolic analog quantities that humans have. We predict that such a model would allow blurrier or imprecise high-magnitude quantifiers. In fact, one could build a model parallel to the one described above by Verguts and colleagues (Verguts & Fias, 2004; section 2.2.2.3) which would learn quantifiers (instead of number symbols) along with nonsymbolic quantities.

An interesting new line of research is the one looking at whether and how quantifiers potentially bias quantity comparison mechanisms. As discussed, there are suggestions and some empirical support for this possibility (Deschamps et al., 2015; Shikhare et al., 2015). Follow-up research could gather more empirical support (in the case of Shikhare and colleagues this was a post-hoc suggestion based on an asymmetrical pattern they observed in their data, which means that a replication is necessary) and compare a wider range of quantifiers (Shikhare and colleagues themselves attempt to do this for 'many' and 'few').

2.3.3 The neuronal perspective

From the neuronal perspective, we are interested in whether processing quantifiers requires the involvement of populations of neurons that also take part in processing nonsymbolic quantities. We have discussed that areas within the prefrontal and parietal cortices, and especially the intraparietal sulcus and area around it, are thought to play a crucial role in both symbolic and nonsymbolic quantity processing. If quantifiers interact with or are references to nonsymbolic quantity representations, as we propose, we would expect these same neuronal populations to be crucial for processing quantifier meaning as well. Alternatively, quantifiers might be represented as a separate, independent network, for example in the left temporal lobe where other semantic categories are thought to be housed according to major theories of language processing by the brain (e.g., Binder & Desai, 2011; Matchin & Hickok, 2020; Ralph, Jefferies, Patterson, & Rogers, 2017).

Given that there are certain differences between quantifiers and nonsymbolic quantity, we do not expect the processing mechanisms to fully overlap, but at least partially. Specifically, for example for deciding which of the possible quantifiers best suits as a description of a cardinality or a proportion, we would expect decision-making processes to be involved. Given the context-sensitivity of quantifiers, the fact that they give rise to pragmatic inferences, and that they need to be read as words before their meaning is understood, we would expect to see some involvement of general language processing areas. Finally, if generalized universal quantifiers rely on logical reasoning, we would expect corresponding mechanisms to also be involved (see e.g., McMillan, Clark, Moore, Devita, & Grossman, 2005; Szymanik, 2016a; Troiani et al., 2009 for discussions of the implications of these differences for the brain regions involved in processing quantifiers). In this review, we only briefly mention studies relevant to these additional processes as discussing them in detail is beyond the scope of the present paper. Instead, we focus on whether the quantity processing system is involved in case of each of the quantifier classes that we have distinguished.

As opposed to research on number symbols, where there is already a long history of neuroimaging research on number symbol processing, there are relatively few studies that look at brain regions subserving quantifier processing (with all of them simply comparing the average amount of brain activity within an area in different conditions[16]). This is why below we review each study in more detail below. For modified numerals specifically, clearly representing the cardinality of a set of objects and its comparison with another set is required before a quantifier judgment or production can happen. For this reason, some of the studies discussed below used the neural correlates of processing modified numerals as a baseline to look at the potential recruitment of the quantity processing system

[16]An exception is a study by Heim and colleagues (2012) that parametrically varied the properties of the stimuli and looked for regions in which activity correlated with this change. This study is discussed further below.

for other quantifiers.

Importantly, as noted, whereas research on number words was mostly focused on representations of number symbols (i.e., what is stored in our cognitive system for each number symbol), in the case of quantifiers questions have also been asked in terms of the processes involved in interpreting them, i.e., about their dynamic evaluation by various mechanisms of the cognitive system.

2.3.3.1 The importance of the parietal cortex: Patient studies

One approach used to look at the neuronal populations important for quantifier processing, parallel to that used for number symbols, is looking at patients with damage in the parietal cortex. Several studies have been conducted with participants with Corticobasal Syndrome (CBS) who are known to have impaired processing of both number symbols and nonsymbolic quantities (as discussed in section 2.2.3.2). Participants were typically asked to make judgments about whether a particular statement or sentence with a quantifier correctly described a picture. For modified numerals, several studies have reported impaired knowledge in CBS patients both relative to healthy age-matched controls and relative to patients with damage to other parts of the brain (McMillan et al., 2006; Morgan et al., 2011; Troiani, Clark, & Grossman, 2011; Troiani et al., 2009). Performance with a limited selection of quantifiers has been investigated in CBS patients. Three studies to date have looked at their performance with the quantifiers 'some' and 'all'. Troiani and colleagues (Troiani et al., 2009) observed worse performance in CBS patients for modified numerals than these quantifiers and interpreted this as evidence for 'some' and 'all' not recruiting parietal areas. However, the CBS patients in fact performed worse for both modified numerals and 'some' and 'all' when compared to the control group of Parkinson's disease patients; thus, it seems like 'some' and 'all' also rely on parietal areas, just to a lesser extent than modified numerals. Supporting this possibility, McMillan and colleagues (McMillan et al., 2006) report worse performance for 'some,' 'all,' and modified numerals combined (this, however, means that only modified numerals might have been responsible for the effect as a group) by CBS patients relative to age-matched controls, patients with Alzheimer's disease, and frontotemporal dementia (which does not typically involve parietal lobe damage). On the other hand, contrary to the results of these two studies, Morgan and colleagues (Morgan et al., 2011) found comparable performance for 'some' and 'all' in CBS patients relative to age-matched controls and frontotemporal dementia patients.

Other quantifiers that have been investigated with CBS patients are 'at least / more / less than half,' again with mixed results. Troiani and colleagues (Troiani et al., 2011) report impaired performance with these quantifierss compared to healthy seniors and a brain-damaged control group, but they were analyzed together with modified numerals, so the latter may have been driving the effect. On the other hand, Morgan and colleagues (Morgan et al., 2011) do not find impaired

performance with these quantifiers.

Finally, one recent study investigated the performance of CBS patients in a production task where they were asked to describe a picture; the authors observed fewer uttered quantifiers by these patients, but did not provide a comparison between different classes of quantifiers (Ash et al., 2016).

Overall, so far it has been consistently observed that CBS patients are impaired in the processing of modified numerals, whereas the results with other quantifier classes remain mixed. For the quantifiers 'at least/more/less than half,' one study observed impaired performance in CBS patients whereas another did not. For 'some' and 'all,' two out of three studies to date suggest that CBS patients are impaired for these quantifiers.

Another line of research with patients has investigated the performance of patients with semantic dementia, a neurodegenerative disorder that mostly affects the left temporal lobe and results in a gradual loss of semantic memory (of semantic concepts such as knowledge about different animals, tools, etc.). Because of this behavioral manifestation of the atrophy, it is thought that the left temporal lobe plays a crucial role in the storage of semantic information. Studies of quantifier knowledge with these patients, therefore, can help us understand whether quantifiers are stored together with these concepts (in which case we expect to see a deterioration in knowledge of quantifiers as well) or separately, e.g., relying more heavily on parietal areas (in this case we expect to see knowledge of quantifiers mostly preserved). In this line of research, so far only two studies with just three patients have been reported. Two out of these three patients had preserved knowledge of the meaning of quantifiers (as well as unimpaired performance in purely quantity-related tasks), while at the same time they had a severely damaged understanding of the meaning of other words (Cappelletti, Butterworth, & Kopelman, 2006; in Cheng et al., 2013 one patient with mild semantic dementia did not have impaired quantifier processing, whereas another patient with severe semantic dementia *was* found to be impaired on quantifier comprehension, although there it could be attributed to a more fundamental deterioration of language skills). These studies tested knowledge of generalized existential and proportional quantifiers, but not generalized universal quantifiers or modified numerals. We can, therefore, exclude the possibility that unimpaired processing was due to the participants being good at mostly or only modified numerals. In addition, we cannot say whether participants were perhaps impaired in terms of their knowledge of generalized universal quantifiers.

Given that different quantifiers may require working memory, logical reasoning, and lexical retrieval, patients with atrophy or damage to other parts of the brain that result in deficits in these capacities have also been found to be impaired regarding some types of quantifiers, but here we refer the reader to the specific studies for more information (see Ash et al., 2016; Morgan et al., 2011, for recent reviews of quantifier processing impairments for other damaged brain areas).

2.3.3.2 The importance of the parietal cortex: Healthy participants

Studies that employed fMRI to investigate brain regions that subserve quantifier processing in healthy adults point to a network of right (or in some studies bilateral) parietal (specifically, the intraparietal sulcus and areas close to it in the inferior and superior parietal cortices) and prefrontal areas, parallel to the network of quantity processing for number symbols and nonsymbolic quantity (Heim et al., 2012, 2016; McMillan et al., 2005; Olm et al., 2014; Troiani et al., 2009). Note that there have been only a few studies to date and they each have relatively few participants, so while we mention the regions in which an increased BOLD signal was observed for quantifiers, these regions are not in fact very specific. For this line of research, it is only possible to make rough generalizations about the involved regions, compared to the line of research into number symbols, where substantially more evidence has been accumulated. The more informative aspect of each of these studies is the comparative involvement of different brain regions in different conditions within the same study.

Several fMRI studies used a sentence-picture verification task similar to that described above (in the context of the interface between quantifiers and nonsymbolic quantity, section 2.3.2.2) to look at brain activity when people verify whether a quantifier correctly describes a visually presented scene. In each trial, participants first saw a sentence containing a quantifier. Subsequently, they saw the same sentence accompanied by a picture depicting a certain number of objects. The participants' task was to indicate whether the sentence was a suitable description of the picture. McMillan and colleagues (McMillan et al., 2005) compared the BOLD response that was present during the display of a sentence together with a picture with the BOLD response that was present when just the sentence was presented. This point in time was thought to reflect the process of verification of the quantifier rather than the reading of the sentence. In this comparison, McMillan and colleagues reported more neural activity in the right inferior parietal cortex for verification than for reading the sentence. The quantifiers they used were 'all,' 'some,' as well as modified numerals, and all were analyzed together as one group. Furthermore, McMillan and colleagues compared activity at this point for different quantifiers. The fact that in all the different quantifier conditions some verification process (and visual array processing) was taking place means that any differences between quantifiers can be attributed to a difference in the verification processes specific to the quantifiers. They did not find more activity in this brain region for modified numerals than for 'some' and 'all,' which speaks to the parietal areas being at least equally active for 'all' and 'some' as for modified numerals.[17] In a different study using a similarly structured sentence-picture verification task, Olm and colleagues (Olm et al., 2014)

[17]Of course, it should be kept in mind that this is a null finding — not seeing a difference is not enough to claim that there was no difference, it only supports such a possibility.

compared the neural activity for verifying sentences with the quantifiers 'some,'[18] 'at least half,'[19] and modified numerals, all analyzed together as a group, with the neural activity for verifying number words (e.g., 'three'; verifying sentences with number words also involves numerical knowledge, so it is an especially strong control ensuring that they captured quantifier meaning processing rather than a pure assessment of the quantity of objects on display[20]). They observed more activity in the bilaterial inferior and superior parietal cortices for quantifiers relative to number words. However, the fact that both of these studies analyzed the modified numerals and other quantifiers together as a group means that the modified numeral alone could be driving the observed effect.

In contrast to the above-described studies, Troiani and colleagues (Troiani et al., 2009) observed more BOLD signal for modified numerals than for 'some' and 'all' in the bilateral intraparietal sulcus, and argue based on this that 'some' and 'all' do not recruit these areas.[21] However, they do not compare brain activity during processing 'some' and 'all' to processing other words, so it is possible that the parietal areas that they identified are involved in processing 'some' and 'all' as well, just to a smaller extent than in processing modified numerals.

Overall, the results of above described fMRI studies are compatible with the possibility that the verification of sentences with the quantifiers 'some,' 'all,' and 'at least half' recruits roughly those brain regions known to be involved in quantity processing.

A different approach to data analysis has been adopted by Heim and colleagues (Heim et al., 2012) who investigated brain activity in response to verifying phrases with proportional quantifiers (specifically, 'many,' 'few,' 'most,' 'very few,' 'more than half,' 'less than half'), also against a visual scene. Instead of comparing brain activity at different points in time, they systematically manipulated the number of target items (blue circles in the case of 'many of the circles are blue') and the ratio between target and comparison items (e.g., the number of blue circles relative to yellow circles). To tap into the semantic processing of quantifiers, they looked for the regions in which activity correlated with the change of ratio between

[18]The article discusses 'logical quantifiers' as a category, but gives only 'some' as an example (with no full stimuli list available), so we are not sure whether there were any other quantifiers included in this category.

[19]Again, this is discussed as a category — 'majority quantifiers' — without giving any other examples, so there might have been more quantifiers under this category.

[20]We already know that parietal lobe areas are involved in estimating and comparing quantity information, so any task that involves this will be bound to result in parietal area activations. The tricky part is finding out what part of such activation is due to quantity processing itself and what part of such activation is due to perhaps the retrieval of and maintenance of the meaning of the quantifier.

[21]This study had a similar design to the two studies described above, except that they presented visual objects whose quantity was to be evaluated against quantifier sentences serially rather than simultaneously as McMillan and colleagues did, and without the sentence with the quantifier present on the display at the same time.

target and comparison items and which were also involved during comprehension of the sentence with the quantifier before that (in this study, sentences were presented auditorily before visual scene presentation). The result of this analysis was thought to specifically reflect evaluation of the meaning of the quantifier and whether it fit the picture. The bilateral intraparietal sulcus and inferior parietal cortex were identified among the regions correlating with the semantic processing of quantifiers. Given that there were no modified numerals among the materials used in this study, these results cannot be attributed to their presence.[22],[23] Thus, for proportional quantifiers we have rather strong evidence for involvement of the brain regions crucial for quantity processing.

2.3.3.3 Single brain region processing quantifiers, symbolic and nonsymbolic quantity?

To date, only one study that we are aware of directly compares fMRI BOLD signal during quantifier comprehension, quantity processing, and comprehension of other words to directly test whether the parietal areas crucial for quantity processing indeed get involved in quantifier processing more than in semantic processing of words in general (Wei et al., 2014). In different trials within this study, participants saw pairs of quantifiers, Arabic digits, number words, dot arrays, frequency adverbs, or animal names and were asked to choose the one that was most similar in meaning to a third stimulus of the same type. Wei and colleagues reasoned that if an overlapping neuronal population subserves the processing of nonsymbolic quantity as well as symbolic quantity in the form of number words, digits, quantifiers (specifically, generalized existential and proportional quantifiers were included), and frequency adverbs (such as 'always,' 'often,' 'never'), this population should be more involved in reasoning about all of these materials when compared to reasoning about animal names (specifically, they expected to find at least one overlapping area when looking at the BOLD signal for each of them as compared to animal names). They did not find any such common area. In contrast, when looking solely at an area that was involved in processing number words, digits, and dot arrays together (i.e., excluding quantifiers and quantity adverbs), they *did* find a region that seemed to participate in processing all of these but not animal names; this area was in the right intraparietal sulcus. This result speaks against quantifier meaning being represented by a neuronal population overlapping with that representing Arabic digits, number words, and dot

[22]One caveat in this study is that the number of tested participants is not reported in the article, meaning that we cannot be sure about how robust/trustworthy these results are.

[23]Interestingly, follow-up fMRI studies pointed specifically to the left inferior frontal gyrus as subserving specifically the semantic/linguistic aspect of quantifier processing/polarity processing (Heim et al., 2012, 2016; see also Wei, Chen, Yang, Zhang, & Zhou, 2014 for corroborating evidence). This is expected given that neuronal populations in this area are considered to be some of the crucial ones for language processing (Friederici, 2002; Hagoort, 2017; Matchin & Hickok, 2020).

array representations. However, the inclusion of frequency adverbs in the analysis together with quantifiers precludes us from drawing strong conclusions since we are not confident that frequency adverbs also involve quantity processing mechanisms. Unfortunately, no analysis excluding frequency adverbs is reported by Wei and colleagues, so further research is needed to make confident conclusions.

2.3.3.4 Questions from the neuronal perspective

Overall, when considering the accumulated evidence from both studies with patients and neuroimaging studies of healthy participants, we see that it is consistent with the possibility that the right or bilateral intraparietal sulcus and surrounding areas are involved in processing modified numerals and various proportional quantifiers. For the generalized existential 'some' and generalized universal 'all' the evidence so far is mixed. The present review makes it immediately clear that only a limited set of quantifiers has in fact been investigated in work with patients and neuroimaging research. Specifically, various generalized existential quantifiers besides 'some' should be investigated[24] (e.g., 'several,' 'a few,' 'a couple'; and the value judgment 'many' and 'few,' if that can be made explicit by the task) as well as generalized universal quantifiers besides 'all' (e.g., 'every,' 'each'). The modified numerals or bare number words could function as a good baseline if the processing of other quantifiers is compared to their processing in the same context. At the same time, these studies should aim to first identify the neuronal populations responsible for nonsymbolic quantity processing in order to allow for drawing conclusions about potential overlap directly within the same study rather than based on previous studies (given that differences in, e.g., data processing procedures or a specific set of participants make it difficult to draw such conclusions based on data across different studies).

An important consideration for future is studies is to be able to distinguish quantifier representations and decision-making or accompanying logical reasoning processes within tasks. The analyses conducted by Heim and colleagues (Heim et al., 2012) address these issues by looking for regions that are involved in *both* processing the sentence with a quantifier without any visual display and processing the visual display itself. A different solution to this issue is to not include a visual quantity comparison task at all, as in the study by Wei and colleagues (Wei et al., 2014); though in their case an additional issue is that the neural populations involved in *reasoning about* quantifiers might not be the same as those representing the meaning of the quantifier.

When thinking about quantifiers in terms of the questions that have been asked for number symbols, one consideration is whether we can expect to see the same format of representations for quantifiers and nonsymbolic quantities. Do quantifiers that are similar in meaning overlap in their representations more than

[24]Note that the above-described study by Wei and colleagues (2014) did include other generalized existential quantifiers, but we believe their analyses are not sufficiently informative.

quantifiers that are far apart in meaning? This question is the same as that raised in section 2.3.2.3 from the behavioral perspective, where we suggested either overlapping or discrete unit representations for quantifiers. Neuroimaging methods can also be used to tap into this question. As discussed, a ratio-dependent similarity has been observed for nonsymbolic quantity representations in adaptation paradigm and RSA analyses of fMRI BOLD data. Using a similar approach for quantifiers, one would expect to see more similarity in neural activation patterns for quantifiers with more overlap. For example, we would expect to see more similar activation patterns for high-magnitude quantifiers than for low-magnitude quantifiers (as discussed above based on the results of Pezzelle and colleagues; (Pezzelle et al., 2018).

In parallel to these studies on number symbols, questions about whether representations of quantifiers and nonsymbolic quantities actually overlap can be investigated using an adaptation paradigm (possibly with a change in notation — adaptation to dot array representing a certain proportion followed by a deviant quantifier that refers to a similar or dissimilar proportion; e.g., 3/10 followed by 'few' or 'many') and RSA analyses (similarly to the study and analysis conducted by Lyons and colleagues (2018) discussed in section 2.2.3.3, one could look at the similarity of activation patterns in response to quantifiers and dot arrays). In a similar vein, one could investigate similarity between 'most' and 'more than half' (this issue is discussed in section 2.3.2.4) by looking at similarity in corresponding neural activations between these quantifiers.

2.4 Summary. Suggested directions of research.

The paper presents an attempt to connect what we know about the relationship between number symbols and nonsymbolic quantity processing to research into the semantics of natural language quantifiers. Both number symbols and natural language quantifiers can be seen as symbolic references to perceptually perceived quantity information. The first part of the paper gave an overview of what we currently know about the relationship between number symbols and nonsymbolic quantity processing and the experimental paradigms that have been used to accumulate that knowledge. In the second part of the paper, we reviewed past studies relevant to the relationship between natural language quantifiers and nonsymbolic quantity processing.

We proposed that generalized existential quantifiers (such as 'several,' 'some,' or 'a couple') can be seen as direct references to approximate cardinality information extracted from nonsymbolic quantity representations, whereas proportional quantifiers (such as 'most,' 'many,' or 'few') can be seen as direct references to extracted ratio information. From behavioral and neuroimaging studies, we know that information about both is computed by the cognitive system representing nonsymbolic quantity. When it comes to generalized universal quantifiers (such as 'all,' 'every,' or 'each'), the connection to the nonsymbolic quantity representations is not necessary, but not clearly excluded either.

Importantly, throughout the second part we presented a number of new research directions and specific questions regarding the processing of quantifiers, which we hope will inspire follow-up research and further theoretical considerations. We now list the most prominent questions:

General:

Q1, section 2.3.0.1 What are natural and robust quantifier classes with respect to symbolic and nonsymbolic quantity processing?

The developmental perspective:

Q2, sections 2.3.1.1, 2.3.1.3 Is children's understanding of quantifiers correlated with their nonsymbolic number acuity? Does improvement in nonsymbolic number acuity result in an improved understanding of quantifiers?

Q3, sections 2.3.1.2, 2.3.1.3 Is the order of quantifier acquisition linked to the development of nonsymbolic number acuity?

Q4, sections 2.3.1.1, 2.3.1.3 Do children make use of nonsymbolic quantity representations to interpret quantifiers against a visual scene? If yes, is the information that children extract from nonsymbolic quantity representations different for each of these quantities? Does the extracted information change over the course of development?

The behavioral perspective:

Q5, sections 2.3.2.2, 2.3.2.5 Do the proportional quantifiers 'most,' 'more/less than,' 'many,' 'few,' and 'more' all extract ratio information from nonsymbolic quantity representations in tasks other than sentence-picture verification?

Q6, sections 2.3.2.2, 2.3.2.5 Do generalized existential quantifiers such as 'some,' 'several,' 'a few,' 'enough,' 'a couple,' 'a dozen' and generalized universal quantifiers such as 'all,' 'every,' 'each' also interface with quantity processing systems, as has been shown for proportional quantifiers? If yes, what kind of information do they extract?

Q7, sections 2.3.2.3, 2.3.2.5 What is the representational format of quantifiers and how does it relate to symbolic and nonsymbolic quantity representations?

Q8, sections 2.3.2.4, 2.3.2.5 Can and do quantifiers bias (i.e., assert top-down influence on) nonsymbolic quantity comparison mechanisms?

The neuronal perspective:

Q9, section 2.3.3.1 Are patients with damage to the parietal lobe impaired in their knowledge of quantifiers other than modified numerals?

Q10, sections 2.3.3.3, 2.3.3.4 Do the neuronal populations involved in processing various classes of quantifiers overlap with the neuronal populations involved in nonsymbolic quantity estimation and comparison?

Q11, sections 2.3.3.1, 2.3.3.4 How is the meaning of quantifiers represented in the brain? Are quantifiers stored together with other semantic concepts (such as animals, tools, etc.) or do they rely on, e.g., parietal lobe areas? Is there more overlap in representations of quantifiers that overlap more in terms of the cardinality or ratio to which they refer? Is there an overlap between specific ratio representations and specific quantifier representations?

Chapter 3

Conducting web-based experiments for numerical cognition research

Abstract

[1]It is becoming increasingly popular and straightforward to collect data in cognitive psychology through web-based studies. In this paper, I review issues around web-based data collection for the purpose of numerical cognition research. Provided that the desired type of data can be collected through a web browser, such online studies offer numerous advantages over traditional forms of physical lab-based data collection, such as gathering data from larger sample sizes in shorter time-windows and easier access to non-local populations. I then present results of two replication studies that employ classical paradigms in numerical cognition research: the number–size congruity paradigm and comparison to a given standard, which also included a masked priming manipulation. In both replications, reaction times and error rates were comparable to original, physical lab-based studies. Consistent with the results of original studies, a distance effect, a congruity effect, and a priming effect were observed. Data collected online thus offers a level of reliability comparable to data collected in a physical lab when it comes to questions in numerical cognition.

3.1 Introduction

Web-based data collection, whereby participants take part in a research study remotely from their own computer, has gained prominence in psychology in the past decade. This is due to its clear advantages: easy access to larger and more diverse samples and speed of data collection (see e.g., Gosling & Mason, 2015; Stewart,

[1]This chapter is based on: Kochari, A. (2019). Conducting Web-based experiments for numerical cognition research. *Journal of Cognition, 2*(1): 39, 1-21. doi: 10.5334/joc.85.

Chandler, & Paolacci, 2017; Woods, Velasco, Levitan, Wan, & Spence, 2015, for reviews). A number of tools have been developed to facilitate the building of experiments with careful timing of experimental stimuli display and accurate response recording within web browsers. This has made web-based data collection suitable for many experimental paradigms typical in cognitive psychology. Likewise, there are now several services offering participant-recruitment from broad participant pools. A number of classical effects in cognitive psychology have been successfully replicated with data collected from participants' web browsers (e.g., Crump, McDonnell, & Gureckis, 2013; Semmelmann & Weigelt, 2017; Zwaan, Pecher, et al., 2018), validating them as viable tools for hypothesis testing. One research area in which this type of data collection can be useful is numerical cognition research, a subfield within cognitive psychology which will be the focus of the present review. Being able to reach a culturally diverse set of participants (e.g., with different traditions of teaching children to count) and to collect data easily in small variations of the same set-up (e.g., multiplication problems with differing levels of difficulty) are advantages that are especially useful in answering questions in numerical cognition.

Technological advancements have allowed for the collection of increasingly sophisticated types of data through participants' web browsers. For a long time web-based studies collected data in the form of questionnaires and survey answers (Birnbaum, 2000; Gosling, Vazire, Srivastava, & John, 2004). Collecting reaction time (RT) data, whereby participants use their own keyboard keys to react to visually or auditorily presented stimuli, has become easier and more prominent as modern browsers (and computers) have become capable of presenting such stimuli with a reasonable timing control (de Leeuw & Motz, 2016; Hilbig, 2016; Semmelmann & Weigelt, 2017). Moreover, with the proliferation of flexible data-collection tools that require less programming knowledge (such as those which are reviewed below), expensive, custom-based software or vast programming skills are no longer necessary for data collection. The most recent developments also allow for tracking the trajectories of participants' mouse movements (Mathur & Reichling, 2018) and recording audio and video (Semmelmann, Hönekopp, & Weigelt, 2017; Semmelmann & Weigelt, 2018). Web-based data collection methods have started being used within numerical cognition research as well; for example, studies looked at numerical information recall performance (Eriksson & Lindskog, 2017), number-line estimation (Landy, Silbert, & Goldin, 2013), mathematical anxiety scores (Cipora, Willmes, Szwarc, & Nuerk, 2018; A. M. Ferguson, Maloney, Fugelsang, & Risko, 2015), reaction time and accuracy in visual stimulus presentation (Cipora, Soltanlou, Reips, & Nuerk, 2019; Gökaydin, Brugger, & Loetscher, 2018; Huber, Nuerk, Reips, & Soltanlou, 2017). Nonetheless, usage of these methods remains rare within numerical cognition[2].

[2]Just 1-3.2% of Google Scholar hits for search terms "numerical cognition", "numerical magnitude", "number processing", "approximate number system" published between 2017-2019

This review focuses on experiments aimed at reaction-time data collection conducted with adult participants (see below, *Collecting good quality data in web-based experiments*, for a short discussion of collecting data with children). The technological requirements for running such experiments are: presentation of instructions and survey questions regarding demographic data, presentation of multiple trials with precise timing, storage of information about accuracy (for a potential analysis of error rates in each condition, as is typical in RT experiments) and timing of button presses by participants. From the point of view of participant commitment, participants should be able to keep their attention on the task, understand and follow instructions, and complete the task as intended (which is something we need to be able to detect). Here, I discuss these issues with a focus on the requirements for posing typical research questions in numerical cognition.

After a short discussion of the advantages and potential problematic aspects of web-based behavioral data collection, I review the available tools for experiment building and participant recruitment. This up-to-date overview will be useful to anyone starting out with web-based data collection, as it aims to answer (or point to where to find answers to) the practical questions surrounding this topic. Thereafter, I present replications of two classical paradigms in numerical cognition research that aim to investigate whether reaction times collected in a web-based study are sensitive enough for typical experimental manipulations in this area. Experiment 1 replicates the size congruity paradigm with numerical size judgment (Besner & Coltheart, 1979; Henik & Tzelgov, 1982). Experiment 2 replicates numerical distance and priming effects in a task where participants were asked to compare a digit to a given standard (Van Opstal et al., 2008). Anticipating the results, in both experiments I successfully replicate the original findings. Note that an earlier web-based study by an independent group successfully replicated a number of effects in a two-digit comparison task (Huber et al., 2017); in yet another web-based study conducted simultaneously with the current project, one more classical effect in numerical cognition - SNARC (Spatial–Numerical Association of Response Codes) effect - was successfully replicated (Cipora et al., 2019). Both current and those replications demonstrate the potential utility of web-based data collection as a tool for research in numerical cognition. In the final section of this manuscript, I offer some advice on how to ensure better data quality in web-based data collection.

3.1.1 Advantages of web-based data collection

One of the advantages of web-based experiments for psychological research is the speed of data collection. Once there is no restriction on the geographical location of the lab, many more participants will usually fit the inclusion criteria of a study.

(as of July 2019) also mention Amazon Mechanical Turk or Prolific Academic (two most popular online participant recruitment services), whereas estimated 11-31% of articles in cognitive and experimental psychology journals in general did so already in 2017 (Stewart et al., 2017)

Thus, more people will be available to participate. There is also no restriction in terms of simultaneously available lab space or computers: each participant completes the study in their own home, and, provided there is no technical limitation from the web server, many participants can complete the same study simultaneously. Lastly, if the researcher makes use of the participant recruitment tools (described below), time is also saved on appointment management: there is no need to schedule each participant and there are no delays related to absentee participants. Speed of data collection not only saves time but also allows for data collection with samples that would not be feasible if the study was not web-based. For example, questions about variability of Approximate Number System in humans as species require extremely large sample sizes (such an investigation of Approximate Number system acuity with 10.000 participants was conducted by Halberda, Ly, Wilmer, Naiman, & Germine, 2012).

Another advantage to this geographical flexibility is that one can reach a population that is not otherwise available or accessible. Web-based data collection makes it easier to recruit non-student samples, non-WEIRD samples (Western, Educated, Industrialized, Rich, Democratic samples; Henrich, Heine, & Norenzayan, 2010), or participants with various linguistic or cultural backgrounds. With regards to the former, we should keep in mind that some level of computer literacy and access to high-speed internet is a necessary prerequisite for participating in web-based studies, so it does not completely solve the issue with WEIRD participants (see e.g., Paolacci & Chandler, 2014; Peer, Brandimarte, Samat, & Acquisti, 2017, for demographic characteristics of online participant pools). The possibility to easily reach participants with different cultural and linguistic backgrounds facilitates the verification of cross-linguistic and cross-cultural experimental effects: one can recruit participants from different populations without conducting data collection at multiple physical locations. In numerical cognition research, such comparisons could be, for example, between cultures which have different customs for teaching children how to count (e.g., Lyle, Wylie, & Morsanyi, 2019; Miller & Stigler, 1987), or from populations which read and write from left to right as opposed to right to left, which would be relevant for a question investigating mental number line (e.g., Pitt & Casasanto, 2016; Shaki, Fischer, & Petrusic, 2009).

Yet another noteworthy advantage of running web-based experiments is the fact that experiments created for web browsers can be more easily shared between researchers. Below, I list some of the tools for programming experiments to run in web browsers. In most cases, the data-collection script will run on any computer with a web browser and can be modified with any text-editing software (although it should be noted that recording the collected data will require a basic web-server or web-server simulator). Unlike many traditional experiment-building tools, there is often no need to pay for licensing. This means that a researcher can simply send the experiment files to a colleague, or upload them as part of the supplemental online materials of a study (as I do for the experiments I present in this paper). Moreover, the same data-collection script can be used both to collect

data remotely from participants' own computers as well as in a physical lab set-up. We now know that findings in psychological research in general suffer from issues of low reproducibility and replicability (Collaboration, 2015; R. A. Klein et al., 2018). Although they have not been investigated specifically within numerical cognition research, these issues are most likely present there as well. Being able to easily share data-collection scripts between different laboratories will allow for close replications of reported effects, improving robustness of findings in the field.

Finally, web-based experiments can be considerably cheaper than lab-based studies if participant recruitment services are used. Contrary to a common belief, however, this is not because the participants are underpaid (in fact, not paying participants a decent amount is an ethically questionable practice; see e.g., Fort, Adda, & Cohen, 2011; Gleibs, 2017, for a discussion of this point), but because of the costs saved on experiment administration. Web-based experiments do not require research assistants to run them, and participant-recruitment services eliminate the need to spend time on scheduling participants or administering payments to each individual participant.

3.1.2 Potential problematic aspects of web-based data collection

As previously mentioned, until recently a skilled web-programmer would have been required to build a reliable web-based experiment, which was problematic. However, various free, intuitive tools built specifically for this purpose are currently available. I give an overview of these tools in the next section.

Whilst participants' environments in traditional lab-experiments are tightly controlled, we have no oversight of participants' environments in web-based experiments. This means that the participants may not be paying as much attention to the task at hand as we may wish (see Chandler, Mueller, & Paolacci, 2014; Necka, Cacioppo, Norman, & Cacioppo, 2016, who found that online participants are often multitasking when participating in studies). Researchers normally explicitly ask participants to be in a quiet room and to pay attention only to the task at hand. However, we have no way to enforce or check for compliance with these instructions. For research questions that claim to investigate everyday brain-function, this may actually prove to be a more realistic experiment environment - for example, when participants are asked to give approximate numerical judgments (as, for example, in Landy et al., 2013). On the other hand, for research questions investigating small effect sizes, an environment filled with distractions may result in noise that conceals the effect. Another possibility is that participants cheat - for example, in a multiplication task without a time restriction they may be solving the given task on a calculator. One can explicitly ask participants if they cheated at the end of the experiment, but we again have no way to know with certainty that they were honest. In the section *Collecting good quality data*

in web-based experiments below, I give some tips on how to maximize participants' attention during the experiment and how to filter out those that did not complete the study honestly. However, since it is impossible to completely avoid these issues, their existence must be taken into account during experiment design and interpretation stages.

The more worrying aspect of collecting reaction time data in web browsers is the accuracy of the stimulus presentation times and of the recorded reaction times. The difficulty with timing of the presentation of visual stimuli is due to varying monitor refresh rates: in order to time a stimulus exactly, one has to specify it in such a way that it takes an exact number of refresh rates (see Elze & Tanner, 2012; Woods et al., 2015, for more detailed discussions). While in a lab set-up one can set the timing based on the known exact refresh rate of the monitor used to run the experiment, it is not possible to do so for web-based experiments as pages loaded to a web browser do not have access to information about the refresh rate of a remote monitor. If a visual stimulus is supposed to appear or disappear at a time that does not coincide with a refresh, it will only do so during the next refresh. In case of auditory stimuli, there will also be a different delay for the different computers and speakers that participants use. However, these timing issues are not as problematic as it may seem, since the delays remain more or less stable within each experimental session, so it will be approximately the same for each trial done by a participant. Thus, it should not be problematic for within-participant designs. This is supported by the results of a study by Reimers and Stewart (2015) who tested stimulus display durations across multiple computers and browsers. They found that stimuli were often presented for around 10-20 ms longer than intended, but within-system variability was small. However, note that experimental designs that go beyond simple visual or auditory stimulus presentation might have unacceptable timing issues; for example, timing lags were found to vary substantially for different browsers and computers when synchronization of auditory and visual stimulus onset was required (Reimers & Stewart, 2016); this issue would, for example, hinder web-based administration of paradigms requiring cross-modal numerical stimulus presentation (as in Lin & Göbel, 2019).

Another problematic aspect is related to delays in reaction time recording: different keyboards will have different delays between the pressing of a key and detection of the press (Neath, Earle, Hallett, & Surprenant, 2011; Plant & Turner, 2009). There will also be delays in RT recording related to inaccuracies in web browsers and to the processing speed of the computer. Multiple studies have compared recorded reaction times in a lab set-up and a web browser-based collection, and they all consistently find delayed RTs for the latter of 25–100 ms (de Leeuw & Motz, 2016; Hilbig, 2016; Reimers & Stewart, 2015; Semmelmann & Weigelt, 2017). Importantly, again, the within-participant variability was stable, and, therefore, the delayed RTs did not affect the size of the observed differences between conditions in within-participant designs (de Leeuw & Motz, 2016; Hilbig,

2016; Reimers & Stewart, 2015; Semmelmann & Weigelt, 2017). When it comes to between-participant designs, the different delays for different participants can potentially be compensated for by testing a larger number of participants in each group (Reimers & Stewart, 2015).

One general issue with using online recruitment services is that participants are likely to complete many studies over time and, therefore, there is a high likelihood that they have experience with similar experimental paradigms or with completing artificial tasks in general. In other words, some of these participants might not be considered naive to the task (Chandler et al., 2014; Peer et al., 2017; Stewart et al., 2015). Participant naivety to the experimental manipulation is often desirable as it is an important assumption of some paradigms (see Chandler, Paolacci, Peer, Mueller, & Ratliff, 2015; Weber & Cook, 1972, for reviews of cases where participant non-naivety can lead to different effect sizes). However, typically, the effects that we are interested in in cognitive psychology, including in numerical cognition research, are robust to participant non-naivety (see Zwaan, Pecher, et al., 2018, for successful replications of classical cognitive psychology effects with non-naive participants). This aspect of web-based data collection is thus less problematic for numerical cognition research than it is for some other research areas.

A number of studies have successfully replicated classical effects in cognitive psychology in web-based studies: Stroop, Flanker, Simon, visual search, attentional blink, serial position, masked priming, associative priming, repetition priming, lexical decision task etc. (Barnhoorn, Haasnoot, Bocanegra, & van Steenbergen, 2015; Crump et al., 2013; Hilbig, 2016; Semmelmann & Weigelt, 2017; Zwaan, Pecher, et al., 2018). As already mentioned, empirical data presented in the current manuscript as well as successful replications of other classical numerical cognition effects (Cipora et al., 2019; Huber et al., 2017) extend their suitability to paradigms typical in numerical cognition research. However, as this section discusses, one should keep in mind that there are certain limitations with web-based data collection: not every lab paradigm will work well running within a web browser or be suitable for completion in an environment with possible distractions. Experiment 2 below replicates one such paradigm that is problematic for web-based data collection - masked priming - for which browser timing inaccuracies for short latencies seem to hinder replication of the effect observed in the lab (Crump et al., 2013, experiment 7). The issue with the environment in which the study is completed remains even in case the web browser is able to execute the experiment flawlessly.

3.2 How to set-up a web-based behavioral experiment

Recently, a number of free community-run and fee-based commercial tools for experiment building have become available, making the creation of web browser based experiments possible with minimal to no web-programming skills at all. Another crucial component in web-based data collection is participant recruitment, which has also became more straightforward with the launch of multiple services specifically aimed to meet this particular need. This section of the paper is intended to be an up-to-date, high-level primer regarding all practical aspects of web-based reaction time data collection. More detailed tutorials are available in the published articles and manuals for each specific tool that I refer to below.

3.2.1 Building behavioral experiments for web browsers

Table 3.1 provides an overview of some of the tools available for building experiments to be run in a web browser.[3] These tools differ in the amount of programming knowledge required, in their pre-programmed functionality, and in whether they are free. They make the task as easy as building experiments to be run in traditional physical lab-spaces (such as PsychoPy, DMDX, E-Prime, Presentation etc.). While other technologies such as Adobe Flash were used in the past, presently JavaScript in combination with HTML5 is the preferred technology as the two are supported by all modern browsers. All of the listed tools support manual scripting using basic JavaScript and HTML code (hence, some programming experience would be required), while some also offer a graphical user interface (hence, no programming experience is needed). Because participant recruitment is done separately from experiment building (although the commercial tools also offer help with participant recruitment), it does not matter which exact tool is used for experiment creation.

For the experiments presented in this paper, I used jsPsych (de Leeuw, 2015); I will briefly describe what experiment building is like with this tool, by way of example. jsPsych is a free, community-built and maintained JavaScript library (i.e., a collection of pre-written functions that can be used by themselves or in addition to other code written in JavaScript) that is optimized for accurate stimulus display and reaction-time data collection. Due to its transparent modular architecture, jsPsych is suitable as an experiment creation tool even for researchers with little to no programming experience. Each experiment script is an HTML page with JavaScript code; it is edited with a text editor and run by opening the same HTML file with a web browser. Within this JavaScript code,

[3]Note that, as this is a fast-growing industry at the moment, many new tools are being released and some existing tools are being discontinued, so I do not attempt to list them all here.

Table 3.1: Overview of some of the available tools for building cognitive psychology experiments to run in web browsers. This list is not comprehensive, as often development is discontinued and new tools frequently appear.

Name	**Website**	**Free**	**Graphical interface**	**Introduction paper**
jsPsych	jspsych.org	yes	no	de Leeuw (2015)
lab.js	lab.js.org	yes	yes	Henninger et al. (2019)
PsychoPy/PsychoJS	github.com/psychopy/psychojs	yes	yes	Peirce et al. (2019)
PsyToolkit	psytoolkit.org	yes	no	Stoet (2017)
Gorilla	gorilla.sc	no	yes	Anwyl-Irvine et al. (2019)
LabVanced	labvanced.com	no	yes	-

an experimenter defines each display (page) of the experiment, the stimuli within these displays, the timing, what the participants are allowed to do (e.g. that they proceed to the next display by pressing only a certain key or that the next display is automatically shown after a certain time), and the order of presentation (or randomization parameters). jsPsych takes care of rendering each of the displays according to the set parameters, recording the answer, and sending the collected data to the determined place for storage (see below for more on this). In the most recent release of jsPsych (at the time of writing this paper), text, images, audio and video can be handled as stimuli by pre-programmed functions in jsPsych; multiple types of survey-question responses, button presses, reaction times of button presses and mouse clicks can be collected. For those new to jsPsych, detailed tutorials and a basic experiment template that can be used as a starting point is available in the supporting documentation for jsPsych (with additional experiment scripts that can be used as templates shared by the community).

3.2.2 Hosting the experiment and storing collected data

Once we figure out the stimulus display and data recording script, the next step is to place it on a web server where participants can access it (i.e., it needs to be hosted somewhere) and arrange for the storage of recorded data. What we would like in the end is a link which participants can follow to take part in the experiment; this link is then given to the recruited participants, for example, through participant recruitment tools. There are multiple options for hosting and data storage. Each of the experiment building tools listed in *Table 3.1* has a help page with detailed suggestions for how to arrange data storage. One way is to host an experiment on a rented or university web hosting service and store the data there; this option requires some basic knowledge of configuring web-servers. There are also independent services for experiment hosting and data storage (e.g. psiTurk - https://psiturk.org/ and JATOS - http://www.jatos.org/, both of which are free and community-run). The commercial experiment building tools that I list above offer to take care of it for you in exchange for a fee.

If you choose to host the experiment on a web server yourself, there are multiple ways to get such a server. Many universities provide personal web-hosting space for their employees that has some basic functionality, which would normally be sufficient for running web-based experiments; that is exactly what was used for the hosting and data collection of the experiments described here. Another option is to rent a hosting space from one of a large variety of companies offering web hosting. In both of these cases the researcher needs to ensure that the server is reliable and that the personal data of participants, if such data is collected, is stored securely as per local requirements. The easiest way to store data collected with jsPsych is as a separate CSV file for each participant. This method requires only that the web-server on which the experiment is hosted supports PHP, which most servers will do by default. The data can be saved, for example, throughout the experiment at the end of each experimental trial. More advanced users can configure data storage in databases such as MySQL.

If it only uses text stimuli (as was the case for the experiments presented here), an experiment built using jsPsych is loaded as a whole before it shows anything to the participant and only connects to the web-server to save the collected data. Thus, there will be no delays related to the internet connection speed of the participant. In case the experiment displays images, audio, or video files, it is also possible to make sure that it only starts after all necessary files are loaded to the computer memory to avoid any delays related to retrieving them from the web-server: jsPsych allows preloading of the media.

3.2.3 Participant recruitment tools

The next step in the process is to recruit participants that will complete the experiment. One way would be to find people willing to take part in the experiment for free, recruiting them, for example, through social media. This would be the best or perhaps the only way forward for those aiming to collect data from thousands of participants, and would also require giving participants some motivation other than a financial incentive to take part (for example, Halberda et al., 2012 collected data by presenting it as a game and offering to give them a score at the end). Here, I focus on another way to recruit participants, namely through crowdsourcing platforms where they come specifically to complete tasks in exchange for a financial reward. This is most suitable for a typical study in numerical cognition, since it is not necessarily interesting enough for participants to just want to do it in their leisure time (sometimes one could think of incentives such as finding out how well one does in comparison to the general population [4]), and would require only dozens or hundreds of participants.

[4]Keep in mind, however, that in this case participants are not going to be specifically setting aside time and ensuring they are in a quiet environment to take part in a scientific experiment; the experiment would rather be a quick entertainment for participants.

The crowdsourcing platform that presumably has the largest pool of participants is Amazon Mechanical Turk (https://www.mturk.com) (e.g., M. Buhrmester, Kwang, & Gosling, 2011; M. D. Buhrmester, Talaifar, & Gosling, 2018). Amazon Mechanical Turk is a marketplace where any sorts of tasks that can be completed remotely, on a computer, are given and taken up by participants. The other presently prominent platform is Prolific Academic (https://prolific.ac/) which is geared specifically towards academic research studies (Palan & Schitter, 2018; Peer et al., 2017). As this is a new industry, a number of other similar platforms appear and close down from time to time (see e.g., Peer et al., 2017; Vakharia & Lease, 2015). Both of the above-mentioned platforms allow for some filtering of eligible participants based on basic demographic data that they fill in: for example, based on age, education level, or native language. Besides the payment to the experiment participant, the researcher pays a fee to the crowdsourcing service.

The data of the experiments reported in the present paper were collected using Prolific Academic (henceforth, simply Prolific), so I will also shortly discuss how this particular service works as an example. On Prolific, the researcher creates a study specifying the participant eligibility criteria, the amount of time the experiment should take, and the amount to be paid, and provides a link that participants should follow to complete the study. A short description of the study is also given to participants, based on which they can decide whether to take part or not. Importantly, the researcher also has an opportunity to include, restrict to, or exclude participants that have taken part in previous studies they have offered. Prolific has a minimum required hourly rate to be paid to participants (£5 at the time of writing this paper), and charges a fee for each of the participants (30% at the time of writing this paper). The researcher also has the possibility to give each individual participant a bonus based on their performance in the study.

Prolific currently has just over 40,000 registered participants, all from OECD countries (people living in other countries are not allowed to register as experiment participants). The participants only have the opportunity to complete a study if they are eligible for it based on their demographic information. If participants choose to take part in a study, they follow the link that is given; Prolific logs the time of the start of the experiment. The experiment is run either in a window with a Prolific heading at the top or in a new window. As a way to confirm that the participant has indeed completed the study, the researcher puts a study-specific link (generated by Prolific) on the last page of the experiment. This link takes the participant back to Prolific, which logs the time of completion. The experimenter has to approve the submission (i.e., verify that the participant undertook the study honestly) before the participant gets paid. After the study is completed, demographic data for participants that took part in it, along with the start and end time for each participant, can be downloaded from Prolific.

3.3 Replications of classical behavioral experiments in numerical cognition

In this section, I present replications of two classical and widely used paradigms in numerical cognition research. Observing the comparable effects in a web-based study and one conducted in a traditional lab-based set-up would support the viability of web-based data collection tools for testing hypotheses in numerical cognition.

Experiment 1 was conducted as part of a different research project and the results are primarily reported in another paper. Here, I only briefly describe it for the purpose of demonstrating the feasibility of getting sufficient quality reaction time data in a web-based experiment. For the same reason, this is not a direct replication of any particular study, but rather a replication of the effects in general. Experiment 2 is conducted as a direct replication of part of a study by van Opstal et al (Van Opstal et al., 2008). Besides the presentation of stimuli and the recording of button-press reaction times, this experiment also includes a subliminal priming manipulation. These replications are only meant as demonstrations of technical possibilities, so I do not offer an interpretation of the effects themselves or their theoretical implications. A successful replication would, however, also demonstrate that these effects, whatever they mean, are robust, since they can be observed in a less controlled environment than traditional physical labs.

The scripts used for data collection, commented data analysis scripts, and all data that were collected are available for inspection and download[5]. Note that these scripts can be easily modified for collecting data in similar studies.[6]

3.3.1 Experiment 1: Size congruity effect

In Experiment 1, I replicate a size congruity effect that was first reported several decades ago (Besner & Coltheart, 1979; Henik & Tzelgov, 1982). Since then, the size congruity paradigm in its original and modified forms has been used to answer numerous questions about number and magnitude perception. In this paradigm, participants are presented with two numbers (e.g., digits, number words etc.) on two sides of the screen and are asked to press a button corresponding to the side of the screen with the numerically larger digit. However, the two numbers that are presented can be of a different physical (font-) size: the numerically larger digit can be physically larger (congruent condition), the numerically larger digit can be physically smaller (incongruent condition), or they can be of equal

[5]See *https://osf.io/dy8kf/*, doi: 10.17605/osf.io/dy8kf.

[6]I also hope that these supplemental materials can serve as an example of how web-based data collection can foster transparency and reproducibility (as discussed above) in numerical cognition research.

physical size (neutral condition). Robust congruity effects are typically observed: people are faster at giving responses and make fewer mistakes in the congruent in comparison to the incongruent condition.

Another variable that is traditionally manipulated in this paradigm is how big the difference between two stimuli is: the numerical difference between the two presented numbers can be large or small (e.g., 2-4 vs. 2-8; I refer to this factor as *numerical distance*) or the physical (font-) size difference between the two presented numbers can be large or small (I refer to this factor as *size distance*). Distance is a relevant factor here since we know that it is more difficult to distinguish values that are closer to each other (e.g., 2-4) than values that are further away from each other (e.g., 2-8) (Moyer & Landauer, 1967). In the size congruity paradigm, numerical and size distance have been found to modulate the congruity effect (see e.g., Cohen Kadosh, Henik, & Rubinsten, 2008; Henik & Tzelgov, 1982; Kaufmann et al., 2005; Pinel, Piazza, Le Bihan, & Dehaene, 2004a; Tzelgov, Meyer, & Henik, 1992). The congruity and distance effects in this paradigm have been interpreted as indicating the automaticity of magnitude processing, since information about the irrelevant dimension modulates performance in the relevant dimension, as well as being used as an argument for the existence of some shared magnitude-processing mechanism (e.g., Cohen Kadosh & Henik, 2006; Cohen Kadosh, Lammertyn, & Izard, 2008; Santens & Verguts, 2011; Tzelgov et al., 1992). As mentioned above, it is not my aim here to address the issue of interpreting the effects themselves. Instead, I focus on whether the basic effect is replicable in a web-based set-up.

In the present experiment, participants judged the numerical value of the presented Arabic digits. I manipulated congruity, numerical distance, and size distance. Based on the results of the classical experiments reporting the size congruity effect (Henik & Tzelgov, 1982), I expected to obtain a main effect of congruity, a main effect of numerical distance (because overall, numbers that are further apart from each other should be easier to judge), as well as an interaction between congruity and physical size distance (because disruption of judgment in the incongruent condition will be stronger when the difference in the irrelevant physical (font-) size dimension is larger).

Method

Participants Given that previous studies were able to detect the size congruity effect with 10-20 participants (e.g., Cohen Kadosh, Henik, & Rubinsten, 2008; Henik & Tzelgov, 1982; Kaufmann et al., 2005), I aimed for a sample size of around 25 participants in this task. Participants were recruited via Prolific.ac. The following inclusion criteria were applied: age 18-25, speaking English as a native language, being born and currently living in the UK. Participants received £1.30 for participation. Participants were excluded from the analyses if they spent less than 10 seconds reading the task instructions or if they gave incorrect

responses in more than 15% of the trials.

Twenty-six participants completed the study in full. Two participants were excluded because they gave incorrect responses in more than 15% of the trials; one further participant was excluded due to reading the instructions for less than 10 seconds. Thus, 23 participants in total were included in the analyses (12 female, 11 male; 3 left-handed; 7 students; average age: 27 [range 19-34]; average time spent on the task: 6:03 minutes [range 4:56-10:47]).

Stimuli Eight digit pairs were included: four pairs had a numerical distance of 2 units ('2-4', '3-5', '5-7', '6-8') and four pairs had a numerical distance of 4 units ('2-6', '3-7', '4-8', '5-9'). Each digit pair was displayed in congruent (the digit in the larger font size is numerically larger) and incongruent (the digit in the larger font size is numerically smaller) conditions. Each digit pair was also displayed in two levels of physical size distance: the font sizes were either 64 pt and 72 pt (*small size distance*) or 55 pt and 72 pt (*large size distance*). Finally, each of the trials was repeated twice, once with the larger number on the left side of the screen and once with the larger number on the right side of the screen. This resulted in 8 (digit pairs) * 2 (congruity levels) * 2 (physical size distance levels) * 2 (sides of the screen) = 64 trials in total. I addition, I included 16 neutral trials (both digits were displayed in font size 64 pt) and 16 filler empty trials, in which participants saw a fixation cross, as in regular trials, but in this case it was followed by a blank screen for 1850 ms. In total participants saw 96 trials. While the neutral condition was present in this experiment, I did not include it in the statistical tests, as assessing whether congruity is driven by facilitation or interference (as in e.g., Cohen Kadosh, Henik, & Rubinsten, 2008) was not my goal.

Procedure The experiment was implemented using jsPsych (de Leeuw, 2015). Prior to the experiment, participants agreed to data collection and filled in a questionnaire asking for basic demographic information. Throughout the experiment, they advanced using the space key or the experiment advanced automatically between experimental trials. Participants were instructed to indicate whether the number on the left or on the right was numerically larger by pressing buttons "Q" or "P" correspondingly. They were asked to do so as quickly as possible. An example was given, and they had a chance to practice making the judgments in 4 practice trials.

Each trial started with a fixation cross ('+') displayed for 150 ms in the middle of the screen. It was followed by a display on the screen where one digit was displayed to the left and another digit to the right of the center. The digits were displayed in Arial font. The digits remained on the screen until the participant gave a response or, if no response was given, for 1850 ms. In case of no response, the experiment automatically advanced to the next trial. The inter-trial interval

was a random number between 700 and 1200 ms.

The experiment was divided into 2 blocks of 48 trials, and the participants had a chance to rest between the blocks. The order of trials was fully randomized. The data for each participant was saved as a separate .csv file on the web-server where the experiment was hosted; this file was updated after each trial.

Results

Participants gave incorrect responses in in 3.7% of trials in total (including trials where no response at all was given). This error rate is within the normal error range for this paradigm (approximately 1-6% based on the studies reviewed above). The general RT level was approximately within the 500-650 ms range, which also falls with the normal range of RTs for this paradigm (e.g., it is somewhat faster than the RTs observed by Henik & Tzelgov, 1982, but somewhat slower than those observed by Cohen Kadosh, Henik, & Rubinsten, 2008).

Only RTs of correctly answered trials were included in the analyses. Prior to the analyses, I excluded all trials in which the reaction time was too short (<250 ms) to have been initiated after processing the target, as well as reaction times shorter or longer than 2 standard deviations from the mean for a given participant for a given condition[7]. This resulted in the exclusion of 8.2% of trials. The resulting RTs, split by congruity and numerical distance, are shown in Figure 3.1a. The same RTs, but this time split by congruity and physical size distance, are shown in Figure 3.1b[8].

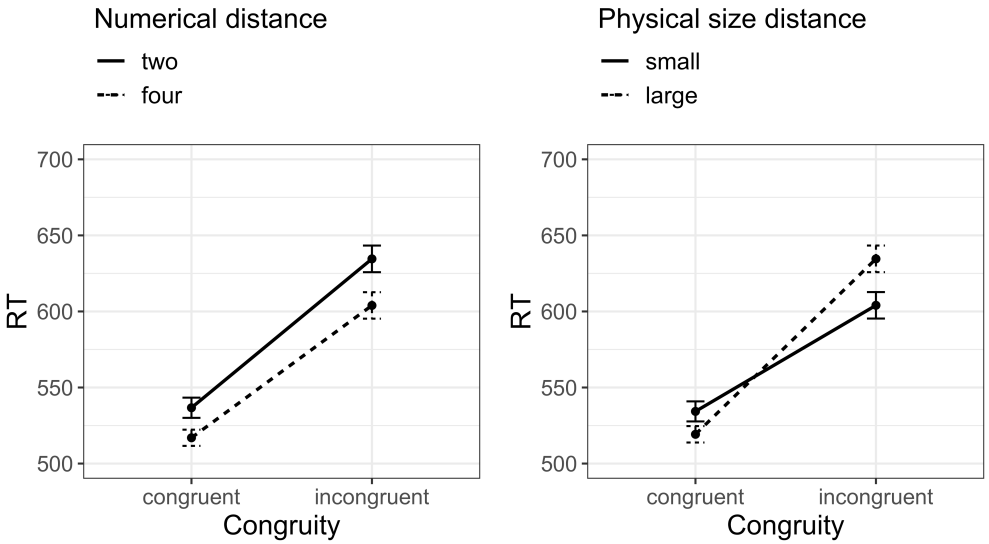
I performed a 2 (congruity: congruent or incongruent) X 2 (numerical distance: two or four units) X 2 (physical size distance: small or large) within-subjects ANOVA on mean correct RTs. All the predictions were borne out by the data. Participants gave faster responses to congruent trials (526 ms) in comparison to incongruent trials (612 ms, difference 86 ms) [$F(1,22) = 60.7$, $p < 0.0001$, $\eta_p^2 = 0.73$]. This effect size is comparable to that observed for the congruity effect in Henik and Tzelgov (difference 116 ms; $\eta_p^2 = 0.9$; Henik & Tzelgov, 1982[9]). I also observed a significant main effect of numerical distance [$F(1,22) = 8.83$, $p = 0.007$, $\eta_p^2 = 0.28$] which was somewhat smaller than the one observed in Henik

[7]Note that Henik and Tzelgov (1982) do not mention any data exclusion, but I did it in this experiment because this experiment was part of a larger project where we applied uniform data processing across all experiments. However, I ran the same analysis of data from Experiment 1 without excluding any reaction times, and the results are the same as the ones presented here.

[8]I opted for depicting the results in two separate plots because a large overlap between the lines within the same plot would hinder visibility of the patterns.

[9]The effect sizes that I report here are based on the reported F-statistic in Henik and Tzelgov (1982), Experiment 2. Note that their analysis also included data from a different task, where participants compared the physical (font-) size of the digits instead of numerosity, so the effect sizes are not based on exactly the same analysis. Note also that their experiment had only 10 participants, so effect sizes estimated in their study have large confidence intervals around them.

Figure 3.1: Mean RTs per congruity in Experiment 1. The error bars depict the standard error value.



(a) Split by numerical distance, collapsing different physical size distances.

(b) Split by physical size distance, collapsing different numerical distances.

and Tzelgov ($\eta_p^2 = 0.68$). Finally, I observed a significant interaction between physical size distance and congruity [$F(1,22) = 15.24$, $p = 0.0007$, $\eta_p^2 = 0.40$; this factor was not manipulated in the study by Henik and Tzelgov]. None of the other effects were significant.

3.3.2 Experiment 2: Numerical distance and priming effects in comparison to standard

In Experiment 2, participants compared a digit to a pre-determined standard (in the present case, the standard was the number 5) and pressed one button in case it was greater than the standard and another button in case it was smaller than the standard. Here too, reaction times and error rates were measured. The theoretical motivation of this paradigm was similar to those that use the size-congruity paradigm (Experiment 1): it is known that the further away the two digits to be compared are, the shorter the reaction times get and the fewer errors participants make; hence, a distance effect should be observed (Dehaene, Dupoux, & Mehler, 1990; Hinrichs, Yurko, & Hu, 1981; Sekuler, Rubin, & Armstrong, 1971, e.g.,). In addition, in this experiment there was a masked prime manipulation - another digit was presented prior to the target digit that participants judged (Dehaene et al., 1998).

Specifically, in this experiment I replicate a study by van Opstal and colleagues

(Van Opstal et al., 2008) that looked at several different effects: the effect of numerical distance between the target digit and the standard digit to which it was to be compared (following their terminology, I will refer to this effect as *comparison distance effect*), the effect of numerical distance between the target digit and the prime digit (following their terminology, I will refer to this effect as *priming distance effect*), and, finally, the *congruity-priming* effect, which refers to the difference between the trials where both the prime and target digit would result in the same response (e.g. both are above standard or both are below standard) and the trials where the prime and the target digit would result in a different response (e.g. the prime is below the standard whereas the target is above the standard). The reasoning of the original study was that unlike the comparison distance effect, which can be explained either by the placement of numbers on an analogue continuum or by response-related processes, the priming effect excludes the response-related processes explanation (see Van Opstal et al., 2008, for the goals of the study). While the original study also used the same paradigm to look at the effects with letters of the alphabet, here I will only replicate their experiment with digits. Van Opstal and colleagues performed the same experiment twice (Experiment 1 and Experiment 2), obtaining the same result; when drawing comparisons to their results below, I will provide the resulting RTs in both of their experiments.

Masked (also referred to as *subliminal*) priming studies require the precise timing of the stimulus display: the prime is usually displayed for some short amount of time, and we need to be sure that the prime has indeed appeared on the monitor and for the specifically defined amount of time which in itself might be an experimental variable. This is challenging in a web-based set-up where we have no control over the exact apparatus that participants are using, and, therefore, cannot synchronize with their monitor refresh rates. One web-based masked priming study by Crump et al (Crump et al., 2013, Exp. 7) attempted to replicate an effect of the compatibility of prime arrows with target arrows (e.g., '>>' primed by '>>' or '<<') in a task where participants simply press a button corresponding to the direction of the arrow. They manipulated the duration of the prime (in 6 steps from 16 to 96 ms) as an experimental factor, expecting the shortest prime durations to result in a negative priming effect (longer RTs after compatible primes) and the longest prime durations to result in a positive priming effect (shorter RTs after compatible primes). They only successfully replicated the priming effects expected for the two longest prime durations, but not the priming effects expected for the shorter prime durations, which were also all trending in the positive direction instead of the expected negative. This was likely due to the fact that with prime durations as short as 16 ms, due to not being synchronized with monitor refresh rates, the primes were sometimes displayed for too long. However, another replication of this effect, which used a different JavaScript library to administer the experiment, did observe the expected positive

and negative priming effects (Barnhoorn et al., 2015, Exp. 3)[10]. Nonetheless, in case the exact duration of the prime display is important for the research question at hand, web-based data collection is not an advised tool since we cannot control it well enough. Web-based data collection would be suitable if it were acceptable for the prime to be displayed for +/– 1 or 2 frames per second longer (which for an average monitor means +/–16 or 32 ms). In the present experiment, the exact duration of the prime was not an experimental factor for the study at hand; moreover, the duration of the masked prime in the study of van Opstal et al was 83 ms - the duration for which both Crump et al and Barnhoorn et al successfully observed priming.

Since this is a direct replication, the present experimental procedure was the same as that described in the van Opstal et al study number task. Whenever I diverged from it, I explicitly mention what exactly was done differently.

Method

Participants Participants were recruited via Prolific, with the same inclusion and exclusion criteria as for Experiment 1, except for one additional inclusion criterion. Namely, in addition participants were not allowed to have completed more than 50 other studies on Prolific. This was done to facilitate participant naivety, which has been raised as a potential issue with participant recruitment through online crowdsourcing services (Chandler et al., 2014; Stewart et al., 2015, see the section *Collecting good quality data in web-based experiments* for a more detailed discussion).

Eighty-one participants completed the experiment across two response button mappings (see below for explanation). They received £ 2.50 for participation. Seven participants were excluded from the analyses due to having given incorrect responses in more than 15% of trials, and two further participants were excluded due to reading the instructions for less than 10 seconds. This resulted in 72 participants being included in the analyses presented below (41 female, 31 male; 13 left-handed; 34 students; average age: 26 [range 18-35]; average time spent on the task: 15:24 minutes [range 11:41-28:22]).

The number of participants for the present study was determined in such a way as to be comparable to that of van Opstal et al. This study included fewer trials than the original study because it is more difficult to ensure participants' attention for longer periods of time in a web-based study (see below, *Collecting good quality data in web-based experiments*, for a discussion). Because there were fewer observations per experimental condition per participant here, I increased the total number of participants in such a way as to end up with approximately the same number of observations per experimental design cell as van Opstal et al.

[10]To my knowledge, these are the only two published studies reporting a subliminal priming task in a web-based set-up.

Stimuli The stimuli in this experiment were the same as in the original van Opstal et al study number task (2008). That is, all numbers from 1 to 9, except 5, functioned as both primes and targets, resulting in 64 different prime-target combinations. However, here the participants saw fewer repetitions of each of the combinations: whereas participants in van Opstal et al saw each prime-target combination 10 times (resulting in 640 experimental trials in total), in the present study participants saw only 4 repetitions of each combination (resulting in 256 experimental trials in total).

Procedure The data were again collected using jsPsych (de Leeuw, 2015). Prior to the experiment, participants agreed to data collection and filled in a questionnaire asking for basic demographic information. Participants were instructed to indicate as quickly as possible whether a number that they would see following '###' was higher or lower than 5. There were two versions of the experiment with different response-button mappings: 38 of the participants included in the analysis were instructed to press 'Q' if the number was lower than 5 and 'P' if the number was higher than 5; 34 of the included participants received the reverse instructions. The presence of prime digits was not mentioned in the instructions. After reading the instructions, participants completed 8 trials as a practice. In these trials, they received feedback about the correctness of the given response immediately after they gave the response. No feedback was provided during the actual experiment.

The stimuli were displayed in the middle of the screen, in white Courier 36 pt font on a black background (Van Opstal and colleagues presented stimuli in font size 32 pt). Each trial started with a fixation cross ('+') displayed for 500 ms. This was followed by a mask ('###') displayed for 100 ms, a prime digit displayed for 83 ms (this would correspond to 5 frames on a monitor with a refresh rate of 60 Hz), and another mask displayed for 100 ms. Finally, the target digit itself was presented until the participant gave a response or for a maximum of 2000 ms. If no response was given, the experiment automatically advanced to the next trial (van Opstal et al did not restrict the time participants had to give a response; I diverged from this in order to make it impossible for the participants to switch their attention to something else during the experiment). The inter-trial interval was 1000 ms.

The experiment was divided into 4 blocks, with the possibility for participants to rest between blocks. In each block, participants saw each of the prime-target combinations once.

Results

Participants gave 3.26% incorrect responses on average (including trials where no response at all was given; van Opstal et al had an error rate of 6.9% in Experiment 1 and 6.5% in Experiment 2). Overall, the reaction times in the present

experiment were ≈90-120 ms longer than in the van Opstal et al data. This is likely due to the fact that the participants in their study completed significantly more trials than in the present study (1280 [640 for number task and 640 for letter task] vs. 256) which meant they were better trained in the task.

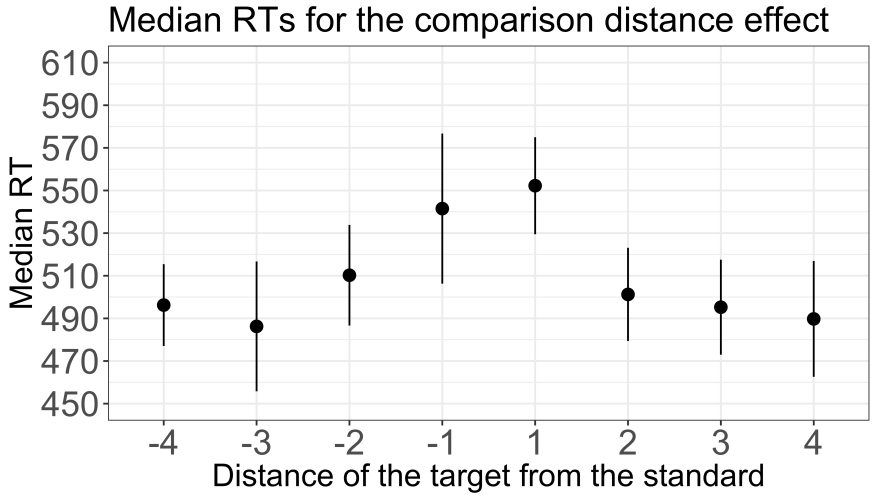
Only the reaction times of correctly answered trials were included in the analysis. Before analysing reaction times, responses that were too fast (<250 ms) to have been initiated after having processed the target digit were excluded; this resulted in the exclusion of 0.05% of trials (van Opstal et al do not report whether they performed an RT cleaning procedure, but I do not consider these RTs meaningful; I did not exclude comparatively long reaction times since the skewed distribution of RTs is likely the reason why Van Opstal et al conducted their analyses on the median RTs). Following van Opstal et al, I also use the median RTs as the dependent variable and performed the same analyses, except that I do not have 'task' as an experimental factor (they had two tasks: number comparison and letter comparison).

Comparison distance effect In order to avoid a confound with the priming distance effect, only trials with identical primes and targets were included in this analysis. Figure 3.2 shows the comparison distance effect. As expected, RTs decreased with the increasing distance of the target digit from the standard. I performed a 2 (size: below/above the standard) X 4 (comparison distance, absolute value: 1, 2, 3, or 4) within-subjects ANOVA. Consistent with the results of van Opstal et al, I observed a significant main effect of comparison distance [$F(3,213) = 10.65$, $p < 0.0001$, $\eta_p^2 = 0.13$]. The size of the observed comparison distance effect was smaller than that reported by van Opstal et al ($\eta_p^2 = 0.38$)[11]. Also consistent with the results of van Opstal et al, there was no main effect of size [$F(1,71) = 2.5$, $p = 0.11$, $\eta_p^2 = 0.03$] and no interaction between comparison distance and size [$F(3,213) = 0.38$, $p = 0.17$, $\eta_p^2 = 0.005$].

Congruity-priming effect In this analysis, I looked at the effect of congruity of the prime and target digit - whether they would result in the same response or in a different response. Here, trials with identical primes and targets were removed to avoid confounding perceptual priming. I performed a 2 (size: below/above the standard) X 2 (congruity) within-subjects ANOVA on median RTs. There were significantly faster reaction times for the congruent (528 ms) in comparison to the incongruent (549 ms, difference 21 ms) prime-target pairs (main effect of congruity: $F(1,71) = 58.44$, $p < 0.0001$, $\eta_p^2 = 0.45$). This is consistent with the results of van Opstal et al (the congruity effect was 26 ms in Experiment 1 and 24 ms in Experiment 2). In addition, the congruity effect was larger for the

[11] The effect sizes that I report here and hereafter are based on the reported F-statistic in van Opstal et al. Note that the analyses in van Opstal et al included a second task, a letter task, so they do not reflect the effect in exactly the same analysis configuration.

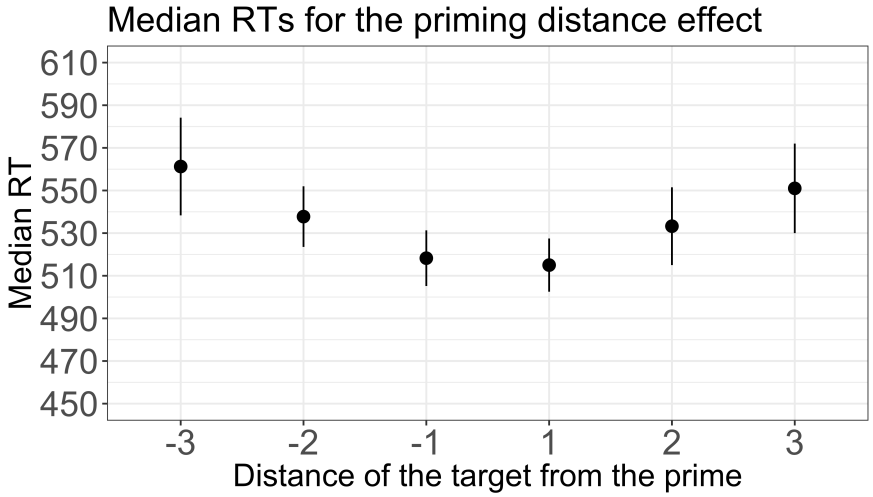
Figure 3.2: Median reaction times (RTs, in milliseconds) for the comparison distance effect. The error bars represent a 95% confidence interval.



digits above the standard (28 ms) than for the digits below the standard (17 ms) (interaction of congruity and size: ($F(1,71) = 4.4$, $p = 0.03$, $\eta_p^2 = 0.05$) which is also consistent with the findings of van Opstal et al (differences: 30 and 22 ms in Experiment 1, 28 and 21 ms in Experiment 2). Finally, in the present study, but not in van Opstal et al, regardless of the congruity, the reaction times were faster overall for the primes below the standard than above the standard; however, this difference was small (difference 5 ms; main effect of size: $F(1,71) = 4.34$, $p = 0.04$, $\eta_p^2 = 0.05$).

Priming distance effect Only congruent trials were included in this analysis. Figure 3.3 shows the priming distance effect. Diverging from the analysis reported by van Opstal et al, I performed this analysis including targets both below and above standard, whereas van Opstal et al. only included targets above the standard (they did so for an independent reason related to the fact that they were interested in comparing priming effects for numbers and letters). I performed a 3 (priming distance, absolute value: 1, 2 or 3) X 2 (size: below/above the standard) within-subjects ANOVA on median RTs. Consistent with the results of van Opstal et al, I observed a main effect of priming distance [$F(2,142) = 28.3$, $p < 0.0001$, $\eta_p^2 = 0.28$]. The size of this effect is comparable to that reported by van Opstal et al ($\eta_p^2 = 0.23$). There was no effect of size [$F(1,71) = 0.85$, $p = 0.3$, $\eta_p^2 = 0.01$] and no interaction between priming distance and size [$F(2,142) = 0.23$, $p = 0.7$, $\eta_p^2 = 0.003$].

Figure 3.3: Median reaction times (RTs, in milliseconds) for the priming distance effect. The error bars represent a 95% confidence interval.



3.3.3 Discussion of the replication results

Both of the presented replications demonstrate that the reaction time effects previously reported in traditional lab studies can be successfully observed when collecting data from participants' web browsers remotely, confirming numerous earlier studies in other sub-fields of cognitive psychology (Crump et al., 2013; Semmelmann & Weigelt, 2017; Zwaan, Pecher, et al., 2018). In Experiment 1, all expected effects in a classical variant of the size congruity paradigm were observed, whereas in Experiment 2 all expected effects in a comparison to the given standard task were observed in a direct replication of van Opstal et al (Van Opstal et al., 2008). The general error rates and reaction times were also within the range of expected values. The results of these experiments convincingly show that getting good quality data is feasible in at least relatively simple web-based numerical cognition experiments.

I draw comparisons between the present results and the results of the original studies based both on statistical significance patterns and on observed effect sizes. However, we should keep in mind that neither of these measures is a good proxy for such comparisons. The statistical significance is a binary, and, therefore, not very informative measure, whereas observed effect sizes are not reliable since both the present and the original studies had low sample sizes and, therefore, do not yield a good estimate of the real effect size.

In Experiment 2, a masked priming manipulation was included in addition to target digit manipulation. Although I do successfully replicate the priming effects observed by van Opstal et al, this is likely because the actual duration of the prime did not matter for this effect. As discussed above, we do not know how long exactly the prime stimuli were displayed since, with JavaScript, there is no

way of finding out exactly how long the stimulus was displayed on the screens of the participants' computers. While in a traditional lab set-up we would be able to set the duration of the stimuli in terms of the number of refresh rates on the monitor used for testing, we cannot do so in this case. Researchers should therefore approach data collection for studies in which the exact duration of the prime matters carefully.

3.4 Collecting good quality data in web-based experiments

While psychologists are already trained to interact with participants in a lab-based setting in such a way as to maximize the quality of the collected data, moving to a web-based set-up introduces a number of new challenges. In this section, I will outline some solutions to common worries associated with web-based data collection.

Ensuring participants have suitable equipment In order to decrease the noise in the collected data due to differences in the equipment used by participants and in order to make sure that the stimuli presentation proceeds in the intended way, we may want to exclude some devices. For example, if the experiment contains audio stimuli, one way to ensure the participants are hearing these stimuli and that they hear them at the intended volume could be to implement a password presented auditorily at the beginning of the experiment, but also repeat it later on to make sure that the equipment stays the same. The same approach can be used if the monitor needs to display certain brightness contrasts and colors (see Woods et al., 2015, for a more detailed discussion).

Ensuring participants are doing the study honestly This is perhaps the most worrisome aspect of web-based data collection for psychologists: participants may simply click through the experiment, respond at random, or give dishonest responses. There a number of simple checks that can be implemented in the experiment. One could use a combination of these checks that suits a study best. If participants respond at random in a straightforward task such as the comparison of numbers, it will be clear from the chance-level performance (for example, two participants in Experiment 1 presented above were excluded for giving incorrect responses in around 50% of trials). If one cannot rely on chance-level performance as an exclusion criterion, it is common to include "catch" trials during the experiment - trials that will unambiguously indicate whether the participant was paying attention (for example, in an experiment where participants need to give their intuitions about the multiplication of 3-digit numbers, one could use the multiplication of single digit numbers as a control; another example would be in-

cluding trials which would say, for example, "Press M" when M is not one of two regular response keys in the experiment). One would then exclude participants who do not reach a certain level of performance in these catch trials regardless of what responses they give in the rest of the experiment.

Even if participants do respond correctly, we still need to make sure that they have followed the instructions precisely (for example, in experiments reported here they should have read that their task is to respond as quickly as possible). One way to make sure this happens is to exclude everyone who read the instructions for less than a certain amount of time that is considered by the researcher to be sufficient[12] (e.g., one participant in Experiment 1 and two participants in Experiment 2 of the present paper were excluded for reading the instructions for less than 10 seconds). Another way is to ask the participants to respond to several quick questions about the instructions before they proceed further in the experiment.[13] Finally, yet another common way to identify dishonest participants is to include a question asking how honest they were at the very end of the experiment, informing them that the response they give will not affect whether they receive payment.

Ensuring participants do not get distracted A common worry is that participants will be multi-tasking during the experiment when we in fact would like them to be focused only on the task at hand (for example, it seems to be common for Amazon Mechanical Turk workers to watch TV or listen to music while doing experiments, see Chandler et al., 2014; Necka et al., 2016). To mitigate this issue, one can strive to administer shorter experiments (for example, no longer than 20 minutes) in order to decrease the chance of participants getting bored and wanting a distraction. This has the consequence that one cannot include many trials and will therefore have fewer datapoints per participant which can potentially be compensated for by collecting data from more participants (for example, this is how I solved this issue for Experiment 2, above; but see Baker et al., 2019; Brysbaert, 2019; Brysbaert & Stevens, 2018 for a discussion of the trade-off between participants and number of experimental trials for statistical power). In addition, there are ways to ensure attention while the experiment is running. For example, I make my experiments auto-paced: trials start and end regardless of whether

[12]Note that by *excluding* I do not mean not paying the participant, but simply excluding them from the analyses. They have contributed their time, and we can never know how well they actually read the instructions, so it is unjustified to withhold payment. On the other hand, I do not pay participants who show only chance performance in straightforward tasks, as in the experiments detailed in the present paper: in those cases, it is clear that they have not even tried to perform the study as instructed.

[13]Both the time spent reading the instructions and the correctness of responses to questions about instructions can be automatically checked by the experiment script while the participant is doing the task, and from there the researcher can decide to either stop the experiment or to give the participant a second chance: for example, by informing them that it is important that they read and understand the instructions and displaying the instructions again.

participants press any buttons (e.g., after 2000 ms of no response, the next trial starts), so the participants do not have an opportunity to divert their attention to something else. Similar to excluding participants if they read the instructions for too short a period, one could also exclude participants if they spend too long on the break between blocks: if someone takes a 10 minute break after 5 minutes of doing the task, they were likely distracted.

Ensuring participant naivety As mentioned, non-naivety is not a large problem for typical cognitive psychology research, and is, therefore, not likely to be a problem for numerical cognition research either (Zwaan, Pecher, et al., 2018). However, one could in principle restrict participation to participants who have completed fewer than a certain number of studies on the participant recruitment service that is used (for example, it is possible to do so on Prolific, and I did so in Experiment 2 above; note, however, that this does not exclude participants who may have completed many studies through another participant recruitment service).

Transparent reporting As is clear from all the possibilities laid out above, there are a multitude of criteria that one can use to exclude certain participants' data from analyses. These researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011) are arguably somewhat larger than in the case of a traditional lab-based data collection, so in the case of web-based experiments it is even more important to preregister the planned exclusion criteria in order to avoid making (conscious or unconscious) biased decisions about data exclusion (Nosek, Ebersole, DeHaven, & Mellor, 2018; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

Web-based data collection with children Everything discussed in this paper so far applies largely to data collected from adult participants. A substantial amount of research in numerical cognition, however, investigates the development of numerical abilities in children, where web-based data collection methods can also be effective in reaching large sample sizes. Unfortunately, the process for collecting developmental data in web-based experiments does not yet seem substantially easier than traditional developmental studies. I will highlight two issues here (for experience with web-based data collection with children and further discussions see Irvine, 2018; Nation, 2018).

Unlike adults, children cannot find and start a study themselves and would not do so in exchange for payment. This means that either a parent or a teacher has to be recruited, start the study and ask the child to complete it. Thus, in this case one still needs to find partner schools that can help with data collection or parents willing to take part in the project. If cooperating teachers are found and helped to go through the process of starting the study and passing it on

to the child, they can repeat it without the researcher's presence later, making this data collection method especially useful for longitudinal studies and studies that can be given to children in multiple school years. Parents could potentially be recruited through social media, in which case the parent should be able to go through the starting-up process and then convinced to not interfere when the child is completing the task. Importantly, one could also administer a home numeracy environment questionnaire to the parent before or after the child completes the task.

There is also an issue with consent, which usually needs to be given by a legal guardian. What is and what is not allowed in this respect will need to be determined by the researcher's local ethical committee. Obtaining consent may not be an issue if a parent starts the experiment and passes it on to the child. For studies administered at school, depending on the nature of a task (whether it falls within a normal set of tasks a child would do at school), level of anonymity of the collected data, and the ethical board stance, one could consider consent as having been given by the teacher when they started the experiment and passed it on to the child.

Despite these additional complications, a number of web-based studies with children are currently being undertaken (e.g., Irvine, 2018; Nation, 2018), including at least one on number cognition (Callaway, 2013).

3.5 Conclusion

In this paper, I have outlined the potential advantages and issues with web browser-based data collection in numerical cognition research. I have also provided pointers for solving practical issues for those starting out with web-based data collection. The successful replications presented here demonstrate that it is not only *possible* to conduct such experiments, but they also yield comparable data quality. Of course, not every type of a study can be conducted in web browsers, but one would be wise to choose this method for studies that *can*, since it saves time and money as well as possibly providing better and larger participant samples. Finally, I have offered some tips for ensuring good data quality. While every study will be unique in the ways in which better data quality can be achieved, by making some adjustments to the ways in which we are trained to ensure data quality, it should be possible to come up with ways to check that participants pay attention, complete the study honestly, etc. for many of the cases.

One final point to address is how we should deal with cases where we will not observe an effect in a web-based study - can we trust it or perhaps it was due to certain timing inaccuracies in web-based data collection? This problem is the same as in case of observing a null result in a lab-based study. The difference is only that in a lab-based study we have presumably eliminated inaccuracies

in timing of stimuli so we are more confident that such a null result is due to the behavior of participants themselves. How do we deal with a null result in a web-based study? In the same way as we would in a lab-based study. For example, one solution would be to design experiments in such a way as to be able to observe a known control effect along with our null-result as a way to make sure the set-up is able to detect an effect of a certain size; another solution would be to move away from null hypothesis significance testing framework towards Bayesian analysis methods that allow to quantify the amount of evidence in favor of the null hypothesis. In general, if a researcher is skeptical about the validity of the results obtained in a web-based study, because they only require a web browser to run, the same experimental scripts can be used both in web-based data collection and in physical lab spaces. One could administer (a part of) the experiment to a smaller sample of participants in a physical lab to verify the obtained result from the web-based study. Overall, a combination of web-based and lab-based data collection methods (verifying the patterns obtained with one by collecting data with another method) would lead to higher confidence in presence of the effect and its generalizability to a larger population.

3.6 Data Accessibility

Raw data, data collection scripts and data analysis scripts for both experiments presented in this manuscript are publicly available: https://osf.io/dy8kf/, DOI 10.17605/OSF.IO/DY8KF.

Chapter 4

Processing symbolic magnitude information conveyed by number words and by scalar adjectives: parallel size congruity and same/different experiments

Abstract

[1]Humans not only process and compare magnitude information such as size, duration, and number perceptually, but they also communicate about these properties using language. In this respect, a relevant class of lexical items are so-called *scalar adjectives* like 'big', 'long', 'loud', etc. which refer to magnitude information. It has been proposed that humans use an amodal and abstract representation format shared by different dimensions, called *the generalized magnitude system* (GMS). In this paper, we test the hypothesis that scalar adjectives are symbolic references to GMS representations, and, therefore, GMS gets involved in processing their meaning. Previously, a parallel hypothesis on the relation between number symbols and GMS representations has been tested with the size congruity paradigm. The results of these experiments showed interference between the processing of number symbols and the processing of physical (font-) size. In the first three experiments of the present study (total N=150), we used the size congruity paradigm and the same/different task to look at the potential interaction between physical size magnitude and numerical magnitude expressed by number words. In the subsequent three experiments (total N=149), we looked at a parallel potential interaction between physical size magnitude and scalar adjective meaning .

[1]This chapter is based on: Kochari, A., & Schriefers, H. Processing symbolic magnitude information conveyed by number words and scalar adjectives: parallel size congruity and same/different experiments. *Manuscript.*

In the size congruity paradigm we observed interference between the processing of the numerical value of number words and the meaning of scalar adjectives on the one hand and physical (font-) size on the other had when participants had to judge the number words or the adjectives (while ignoring physical size). No interference was obtained for the reverse situation, i.e. when participants judged the physical font size (while ignoring numerical value or meaning). The results of the same/different task for both number words and scalar adjectives strongly suggested that the interference that was observed in the size congruity paradigm was likely due to a response conflict at the decision stage of processing rather than due to the recruitment of GMS representations. Taken together, it can be concluded that the size congruity paradigm does not provide evidence in support the hypothesis that GMS representations are used in the processing of number words or scalar adjectives.

4.1 Introduction

A lot of research has been devoted to the question how the human cognitive system estimates and compares magnitudes such as size, length, quantity, loudness, duration, etc. from the perceptual input. Based on the accumulated evidence, some researchers postulated the existence of a generalized analog magnitude system (GMS; Cantlon, Platt, & Brannon, 2009; Lourenco, 2015; Walsh, 2003). Under one such proposal, after our perceptual system takes up the information from the outside world (e.g., in visual modality about length, size or quantity), this information is transformed into an abstract, amodal representation format in which the comparison of magnitudes can be performed. For example, when comparing the lengths of lines, information about length will be mapped onto the same cognitive system as the one used for comparing quantities of objects in two sets, the duration of two auditory signals, etc.

The ability to estimate and compare magnitudes is not only used to perceptually navigate in the world around us, but information about such magnitudes is also communicated to others. At the same time, when others communicate information about magnitudes to us, we need to interpret this information. In this respect, an interesting class of lexical items are so-called *scalar adjectives* (also referred to as *vague* or *gradable*) such as 'tall', 'short', 'long', 'big', 'loud', etc. Interestingly, the meaning of all these adjectives shares a set of features with the format in which magnitudes are thought to be represented in GMS. In this paper, we therefore put forward and test the hypothesis that scalar adjectives are symbolic references to GMS representations, and, thus, when processing their meaning, our language comprehension system makes use of GMS representations.

We present a set of six experiments of which four make use of the *size con-*

gruity paradigm and two are follow-up studies where a *same/different* task is administered. The goal of the experiments was to look at the involvement of GMS in the processing of number words (Experiments 1a-c) and scalar adjectives (Experiments 2a-c) in parallel designs.

The size congruity paradigm (Besner & Coltheart, 1979; Henik & Tzelgov, 1982) is a classical set-up which demonstrates interference between symbolically presented numerical magnitude information (e.g., in the form of Arabic digits) and perceptually presented physical size magnitude information (in the form of the font size of the digits). Congruity effects obtained in this paradigm have been used to argue in favor of shared representations of magnitudes of different dimensions (in the case of Arabic digits, of numerical magnitude and of physical size magnitude). We start by exploring the effects that can be observed with this paradigm, however with with number words instead of digits (Experiments 1a, 1b). Using the same reasoning, in our experiments on scalar adjectives (Experiments 2a, 2b) we ask whether we can observe a parallel congruity effect between the magnitude information expressed by the meaning of scalar adjectives and simultaneously presented physical size magnitude (i.e., the font size in which the adjective are presented). Anticipating on the results, we observed a size congruity effect in the numerical comparison task with number words (Experiment 1a) and meaning comparison with scalar adjectives (Experiment 2a). This congruity effect in principle supports our hypothesis that processing scalar adjectives recruits GMS representations. However, as in the original paradigm, our set-up also allows for an alternative explanation of the congruity effect. This alternative explanation suggests that the congruity effect could arise due to a conflict at the response selection stage of processing rather than due to shared underlying representations. In order to disentangle these two accounts of the congruity effect, we collected data in a novel same/different task. The representational overlap and the conflict at the response selection stage accounts predicted a distinct pattern of effects in this task, allowing us to decide on the locus of the congruity effect (central shared representations versus peripheral competition during response selection). The same/different task was administered for number words (Experiment 1c) as well as for scalar adjectives (Experiment 2c).

4.1.1 Generalized analog magnitude representation system

Let us start with a note on terminology: we will use the word *magnitude* to refer to values along any continuous dimension (e.g., size magnitude, length magnitude, numerical magnitude), and we will use the word *quantity* to refer specifically to numerical magnitude (i.e., the number of distinct individual elements). Furthermore, we will use the word *nonsymbolic magnitude* to refer to magnitudes extracted from perceptual input (e.g., a visually presented array of dots or an auditory sequence of tones) as opposed to *symbolic* references magnitude such as number symbols and scalar adjectives (discussed in the remaining sections of the

introduction).

Most of the work on magnitude processing has been done on numerical magnitude processing, so we start by introducing what is known about numerical magnitude processing. When receiving and evaluating numerical magnitude information from perceptual input, our cognitive system makes an approximation of magnitude (rather than providing us with a precise value[2]), and it has a limited sensitivity with which it can do so. For example, when extracting a quantity from a visual scene, we are able to successfully distinguish a set of 15 dots from a set of 30 dots, but not 28 dots from 30 dots. Performance with nonsymbolic quantities in terms of accuracy and reaction times is dependent on the ratio between the two quantities to be compared such that larger ratios (i.e., larger relative difference in magnitude) lead to faster and more accurate responses than smaller ratios (i.e., smaller relative difference in magnitude; e.g., Buckley & Gillman, 1974; Feigenson et al., 2004; Halberda & Feigenson, 2008; Pica et al., 2004; this is consistent with Weber's law, see e.g., Bar et al., 2019 for a recent discussion). Such performance has been suggested to reflect the operation of the so-called Approximate Number System (ANS) which is thought to be an evolutionary old system shared with other animals (Barth et al., 2003; Dehaene, 1997; Feigenson et al., 2004; Gallistel & Gelman, 2000; Halberda & Feigenson, 2008). Numerical magnitude representations in ANS are thought to be continuous (or *analog*) distributions around a point (similar to a Gaussian distribution) which overlap with neighboring distributions. Two alternative accounts have been proposed[3] - either the spread of the distributions around points increases with increasing quantities or the spread of distributions is same for different quantities but the quantities are logarithmically compressed (e.g., Bar et al., 2019; Dehaene et al., 2008; Feigenson et al., 2004; Gallistel & Gelman, 1992; Merten & Nieder, 2008; Nieder, 2016; we leave this discussion aside as it does not matter for the purpose of the present manuscript).

In the same way as we can approximate a quantity from perceptual input, we can also make such approximations on the length of a line, the duration of an event, the size of an object, etc. These approximate judgments are also limited in precision and, interestingly, are also ratio-dependent (see e.g., *Table 1* in Cohen Kadosh, Lammertyn, & Izard, 2008, for examples). It has been suggested that there is a single shared underlying system for representing and processing perceived magnitudes in various continuous dimensions (including numerical magnitude). This system has been referred to as *generalized (analog) magnitude system* (*GMS*) (Gallistel & Gelman, 2000; Walsh, 2003, 2015) and is usually conceived of as a generalized version of ANS. Magnitudes in such a system are

[2] An exception to this are quantities in the so-called *subitizing* range - up to 3 or 4 (Kaufman et al., 1949; Revkin et al., 2008; Trick & Pylyshyn, 1994).

[3] It should be noted that the existence of a separate quantity encoding ANS has recently been a subject of debate. The alternative to existence of ANS is that information about magnitude of other dimensions is used to infer quantity (Anobile, Cicchini, & Burr, 2016; Gebuis, Cohen Kadosh, & Gevers, 2016; Leibovich, Katzin, Harel, & Henik, 2017).

assumed to be represented in the same way as numerical magnitudes are represented in the ANS - as continuous (analog) distributions around a point which overlap with neighboring values and have increasing uncertainty with increasing values.

There is a number of reasons to postulate the existence of GMS in addition to the parallel ratio-based performance across different dimensions mentioned above. Specifically, a considerable amount of evidence shows transfer or interference effects between magnitude information in different dimensions (e.g., size and quantity, duration and length) when magnitudes in two dimensions have to be judged consecutively or when they are presented simultaneously and one dimension is task-irrelevant (e.g., Bonn & Cantlon, 2017; Casasanto & Boroditsky, 2008; Dormal, Seron, & Pesenti, 2006; Droit-Volet, 2010; Krause, Bekkering, Pratt, & Lindemann, 2017; Lourenco et al., 2016; Möhring, Ramsook, Hirsh-Pasek, Golinkoff, & Newcombe, 2016; Oliveri et al., 2008; Stevens, Mack, & Stevens, 1960; Xuan et al., 2007). For example, in a recent study Bonn and Cantlon (2017) observed spontaneous extraction and transfer of ratio information between e.g., size and duration or size and loudness. Specifically, when asked to judge the similarity of sequences of stimuli in two different dimensions, participants rated stimulus sequences which preserved the ratio information across dimensions as more similar than sequences which preserved only rank information across dimensions. It has also been claimed that there is a similar developmental pattern in the precision with which children are able to discriminate magnitudes in different dimensions (Feigenson, 2007). Finally, there is neuropsychological and neuroimaging evidence suggesting that potentially overlapping neural populations in the intraparietal cortex are involved in the processing magnitudes in different dimensions (e.g., Chassy & Grodd, 2012; Fias, Lammertyn, Reynvoet, Dupont, & a Orban, 2003; Pinel, Piazza, Le Bihan, & Dehaene, 2004b; Sokolowski, Fias, Bosah Ononye, & Ansari, 2017; Zorzi, Priftis, & Umiltà, 2002; see also Nieder, 2016 for references to studies which observe individual neurons responsive to magnitudes in different dimensions).

It should be noted, however, that there is also some debate about the existence of the GMS and about what exactly is shared between dimensions (for reviews and discussion, see Bonn & Cantlon, 2012; Cantlon et al., 2009; Cohen Kadosh, Lammertyn, & Izard, 2008; Leibovich et al., 2017; Lourenco, 2015; Sokolowski, Fias, Bosah Ononye, & Ansari, 2017; Van Opstal & Verguts, 2013; Yates et al., 2012). Evidence against a shared GMS comes, for example, from the observation that transfer effects between magnitudes are not always bidirectional (e.g., Bonn, 2015; Merritt, Casasanto, & Brannon, 2010; Roitman, Brannon, Andrews, & Platt, 2007). Also, the results of interference studies and neuroimaging studies could have an alternative interpretation in terms of learned associations due to co-occurrence in natural environments (see Bonn, 2015; de Hevia & Spelke, 2009; van Galen & Reitsma, 2008).

In the present research project, we assume that there exists a GMS-like mech-

anism that is responsible for computing relative magnitude in various dimensions. This shared mechanism computes relative magnitudes when we are comparing, for example, numerical magnitudes, size magnitudes, length magnitudes, duration magnitudes, etc. from *perceptual* input.

4.1.2 Processing number symbols

Our cognitive system can receive and process quantity information not only from perceptual input, but also *symbolically* - e.g., using Arabic digits ('3','5') or number words ('three', 'five'). Note that number symbols refer to exact, discrete quantities whereas nonsymbolic quantity is perceived in a continuous format, i.e., without sharp boundaries (e.g., Bar et al., 2019; Leibovich, Diesendruck, Rubinsten, & Henik, 2013).

To what extent number symbols (e.g., an Arabic digit) are represented by the same cognitive system or recruit the same processing mechanisms as perceptual, nonsymbolic quantity remains a matter of a debate (see e.g., Nieder, 2016; Piazza & Eger, 2016; Sokolowski & Ansari, 2016; Wilkey & Ansari, 2019, for extensive reviews). It has been suggested that — as a cultural invention — number symbols use (or *recycle*) the evolutionary older ANS-type of neural representations of quantity (Dehaene & Cohen, 2007). To explore this possibility, one line of research looked at whether parallel behavioral effects can be observed for both number symbols and nonsymbolic quantity which would suggest shared representations or at least shared processing mechanisms. Parallel ratio-based performance with both symbolic and nonsymbolic numerical magnitudes has been reported, for example, in quantity comparison tasks (e.g., Dehaene, 2007; Dehaene & Akhavein, 1995; Moyer & Landauer, 1967), though the interpretation of these effects as supporting shared cognitive systems for the two formats has been contested (e.g., Kojouharova & Krajcsi, 2018; Krajcsi, Lengyel, & Kojouharova, 2018; Verguts et al., 2005). Further evidence comes from matching and priming paradigms that showed that number symbols closer to each other in terms of their numerical magnitude seem to have more overlap in their representations than number symbols further away from each other, as would be expected if they recruit the nonsymbolic, ANS representations (e.g., Defever et al., 2011; Reynvoet, De Smedt, & Van den Bussche, 2009; Sasanguie et al., 2011; Van Opstal et al., 2008; Van Opstal & Verguts, 2011), but there is again some counter-evidence (Roggeman et al., 2007; Sasanguie et al., 2017).

Brain imaging studies observe neuronal activity associated with processing symbolically and nonsymbolically presented quantities in partially overlapping regions of parietal and frontal lobes (see e.g., Sokolowski & Ansari, 2016; Sokolowski, Fias, Mousa, & Ansari, 2017, for reviews). Moreover, a number of studies reports ratio-dependent changes in the amount of BOLD signal in the intraparietal cortex when processing symbolic and nonsymbolic quantities (e.g., Cantlon et al., 2006; Goffin et al., 2019; He et al., 2015; Holloway et al., 2012; Piazza et al.,

2004; Vogel et al., 2017). Piazza and colleagues (Piazza et al., 2007) presented evidence for such ratio-dependent changes within a single region across symbolic and nonsymbolic quantity presentation which supports the suggestion that the neuronal populations are at least partially shared for them. On the other hand, recent studies making use of activation pattern analysis techniques (representational similarity analysis [RSA] and multivoxel classification) report that the pattern of voxelwise activity correlations in the intraparietal cortex and some other areas corresponded to overlapping analog representations for nonsymbolically presented quantities, but not for number symbols (Bulthé et al., 2014, 2015; Lyons et al., 2015; Lyons & Beilock, 2018). In addition, these studies did not find a region that would encode a certain quantity presented in both symbolic and nonsymbolic formats (e.g., '6' and six dots). Both pieces of evidence speak against shared neuronal representations of symbolic and nonsymbolic quantity. Nonetheless, evidence from these latter studies does not completely exclude the possibility of shared representations since this conclusion is based on a null result and since there are certain concerns regarding the sensitivity of these analyses (see Eger, 2016; Wilkey & Ansari, 2019 for arguments). Moreover, neural network modeling shows that even when number symbols are represented by the same neurons as nonsymbolic quantities, they only partially recycle the representations of nonsymbolic quantities and are more efficient in encoding, meaning that they do not necessarily have to show the same pattern of activation (Verguts & Fias, 2004).

Given the mixed evidence, deciding whether and to what extent symbolic and nonsymbolic quantity have overlapping representations requires further research. The present project does not aim to resolve this issue. Rather, the central goal of the present project is to apply an existing paradigm that has previously been used to investigate the potential relationship between number symbols and nonsymbolic quantity as a starting point for investigation of processing scalar adjectives (introduced in the next subsection). However, in doing so, as a first step we also use this experimental paradigm to number words. Thus, the results of the present study will partially also contribute to the number symbol processing research.

Recall that in the previous section we discussed the proposal that there exists a single shared cognitive system for processing magnitudes from perceptual input along various dimensions, GMS. Assuming a GMS-like cognitive system exists, if number symbol representations are indeed partially shared with nonsymbolic numerical magnitude representations, then number symbol representations should also be partially shared with nonsymbolic magnitude representations in other dimensions. A number of studies has presented evidence for this relationship in the past. As we discuss below in detail, this has been more convincingly demonstrated in case of Arabic digits than in case of number word processing. In the present study, we test this prediction in case of specifically number words as references to numerical magnitude and size magnitude (Experiments 1a-1c). Before we do that, let us introduce the parallel hypothesis regarding scalar adjectives, i.e., the

type of magnitude information carrying elements that we are primarily interested in in this project.

4.1.3 Scalar adjectives

Similarly to numerical magnitude, magnitude information along other dimensions can also be conveyed symbolically. In natural language, we can describe an object's magnitude along a particular dimension using adjectives such as 'big' and 'small', 'long' and 'short', 'loud' and 'quiet', etc. For example, we can describe a new TV at our neighbors' house as being 'big' or this morning's weekly work meeting as lasting 'long'. This class of adjectives is referred to as *scalar adjectives* (or sometimes *vague* or *gradable adjectives*; see e.g., Frazier, Clifton, & Stolterfoht, 2008; Solt, 2015b; van Rooij, 2011b for reviews).

Scalar adjectives seem to possess some of the properties of the GMS representation format (earlier observed by Fults, 2011). First, we can use adjectives like 'tall' to describe quite different heights - e.g., that of buildings, trees or people. This suggests that these adjectives are flexible in their magnitude reference and what seems to matter for applicability of these adjectives is relative magnitude in a given context, not the absolute value. This property is consistent with our suggestion that they are referring to GMS-like representations because there too, what matters in comparison are relative rather than absolute values.

The second relevant property is that these adjectives lack sharp boundaries that determine when they do and when they do not apply as descriptions of a particular magnitude. For example, there is no one specific height that we refer to as being 'tall', not even if we talk about something specific, e.g., 'a tall building'. Furthermore, if we earlier referred to some building as being 'tall' and now see a different building that is only slightly shorter, then we would have to admit that 'tall' also applies to this slightly shorter building, and in this situation it is impossible to come up with a strict criterion for when a building is not tall anymore when we take small steps (relative to the absolute magnitude). Thus these symbolic magnitudes are like nonsymbolic magnitudes; they are represented in a continuous format with no strict boundaries and small differences between magnitudes are not perceptible. Relatedly, we can even count the number of floors of a building or measure the size of an object using exact numbers, but we would still not know when 'tall' and 'big' exactly do and do not apply. This once again demonstrates that these adjectives do not refer to or involve discrete magnitudes in interpretation.

Scalar adjectives have been a subject of extensive research within philosophy of language and semantics, but received relatively little attention in psycholinguistics. Researchers in psycholinguistics may be familiar with scalar adjectives from the now classical work of Sedivy and colleagues (Sedivy, Tanenhaus, Chambers, & Carlson, 1999) who were interested in whether participants will take into account the meaning of the scalar adjectives immediately as they hear them (i.e., incre-

mentally) and interpret them in relation to the contrast between objects that they simultaneously see on the display. Other studies on scalar adjectives were follow-ups to this study, further exploring online interpretation of scalar adjectives in a given context (e.g., Aparicio, Xiang, & Kennedy, 2016; Rubio-Fernandez, Terrasa, Shukla, & Jara-Ettinger, 2019; Wolter, Gorman, & Tanenhaus, 2011). A different research line used the fact that the meaning of scalar adjectives depends on the noun that it combines with to investigate timing and neural correlates of semantic composition of minimal adjective-noun phrases (specifically, they collected magnetoencephalography data to compare processing of e.g., 'large table' vs. 'wooden table', but did not observe conclusive evidence for presence of any differences in the composition process; Kochari, Lewis, Schoffelen, & Schriefers, 2020; Ziegler & Pylkkänen, 2016).

Let us now turn to the present research project. Above, we discussed the generalized magnitude system that we assume is recruited for processing nonsymbolic magnitudes in different dimensions. Furthermore, we proposed that scalar adjectives in language can be seen as symbolic references to the magnitudes that they refer to, and that they do so. Departing from these two observations, in the present study we ask whether such GMS-like representations are recruited in the processing of scalar adjectives just as it has been shown by a number of studies on number symbols and nonsymbolic numerical magnitude and on other non-numerical magnitudes. We suggest that our language processing system makes use of the generalized magnitude system representations during the retrieval of the meaning of scalar adjectives and the construction of a mental model of the communicated information. Thus, for example, in order to understand a phrase like 'a long meeting', we make use of the GMS in order to imagine this meeting being longer than some other meeting we experienced. We test this hypothesis by investigating whether we can observe an interference between magnitude information conveyed perceptually and magnitude information extracted when processing scalar adjectives. We now turn to the discussion of the experimental paradigm that we use.

4.1.4 Size congruity effect as an indicator of shared representations across different magnitude dimensions

A classical experimental set-up that has been used to demonstrate interference of symbolic and nonsymbolic magnitude information from different dimensions is the number size congruity paradigm (Besner & Coltheart, 1979; Henik & Tzelgov, 1982; for recent studies using this paradigm see e.g., Arend & Henik, 2015; Cohen Kadosh, Henik, & Rubinsten, 2008; Gabay, Leibovich, Henik, & Gronau, 2013; Leibovich et al., 2013; Santens & Verguts, 2011). In this paradigm which is similar to the Stroop task, participants are typically presented with two Arabic digits side by side on a screen. They are asked to decide which one has the

larger or the smaller numerical magnitude. The relative font size of the two digits (which is irrelevant for the task) is manipulated such that it either agrees with the numerical magnitude information (**5** 3; congruent condition) or is in conflict with it (5 **3**, incongruent condition). In the second version of the paradigm, the dimensions are reversed - participants have to ignore the numerical magnitude and instead decide which of the two presented digits is of physically larger or smaller font size. In this version of the task, again, numerical information either agrees or is in conflict with size information. A robust congruity effect has been observed in both versions: reaction times are shorter in the congruent condition than in the incongruent condition.

Interestingly, in addition to the congruity effect, an interaction with the magnitude distance information in the to-be-ignored dimension can be observed. Specifically, the size of the congruity effect is modulated by the extent to which two digits differ in the task-irrelevant dimension in terms of the numerical magnitude or in terms of physical font size (referred to as *size* or *numerical distance*; e.g., Arend & Henik, 2015; Cohen Kadosh, Henik, & Rubinsten, 2008; Henik & Tzelgov, 1982; Kaufmann et al., 2005).[4] Specifically, with a larger difference between the size or numerical magnitude in the task-irrelevant dimension one observes a larger congruity effect (i.e., more interference). This finding shows that the congruity effect is not just driven by the presence of *some* conflicting information, but rather the size or strength of this conflicting information matters.

The size congruity effect has been interpreted as evidence for two aspects of magnitude processing: automaticity of computation of numerical and physical size magnitude (e.g., Dadon & Henik, 2017; Henik & Tzelgov, 1982; Pansky & Algom, 1999; Tzelgov et al., 1992) and shared representations underlying numerical and size magnitudes (e.g., Arend & Henik, 2015; Cohen Kadosh, Lammertyn, & Izard, 2008; Schwarz & Heinze, 1998). Let us consider the first point. The size congruity effect shows that both physical and numerical magnitude are able to interfere with performance even though they are task-irrelevant. Because information in the task-irrelevant dimension could not be completely ignored in this task, it has been suggested that physical size and numerical magnitude are automatically computed (in case of physical size) or retrieved (in case of numerical magnitude). To what extent are these computations automatic? On strong automaticity account, no general processing resources would be required for computation or retrieval of magnitude. However, the congruity effect has been shown to be modulated (but not eliminated) by the discriminability of physical sizes and digit pairs as well as to some extent by practice and motivation, so strong automaticity can be ruled out (Algom, Dekel, & Pansky, 1996; Dadon & Henik, 2017; Pansky & Algom, 1999). Instead, the size and numerical magnitude computations seem to be automatic in the sense that activation of magnitude representations is obligatory (at least in

[4]Not that the size congruity effect is also modulated by the difference between magnitudes in the task-relevant dimension, but this is not relevant for the present argument.

the size congruity paradigm), but does require processing resources, and cognitive control can be exerted to some extent (Dadon & Henik, 2017; Pansky & Algom, 1999).[5] Our main goal in this study is to investigate whether there is a magnitude representation shared by physical size and scalar adjectives, but automaticity will be discussed to some extent as well as it is a prerequisite for the congruity effect in the size congruity paradigm.

The size congruity effect has also been interpreted as evidence in favor of shared representations of numerical magnitude and physical size magnitude (but see Risko, Maloney, & Fugelsang, 2013; Santens & Verguts, 2011 for alternative interpretations, to be discussed below). Specifically, it has been proposed that both the retrieved numerical magnitude of a digit and its size magnitude are encoded into a common GMS representation, and that the congruity effect occurs due to a conflict or a match at this encoding stage (e.g., Arend & Henik, 2015; Cohen Kadosh, Lammertyn, & Izard, 2008; Reike & Schwarz, 2017; Schwarz & Heinze, 1998; Szucs & Soltesz, 2008). In addition, because in this paradigm the numerical magnitude is presented symbolically whereas the size magnitude is perceptual, the observed congruity effect also supports the claim that number symbols make use of at least partially shared representations not only with perceptual numerical magnitude, but also perceptual magnitude in other dimensions. Using this paradigm, congruity effects have also been observed with other dimensions - for example, number and area (Hurewitz, Gelman, & Schnitzer, 2006), and number and luminance (Cohen Kadosh & Henik, 2006; Pinel et al., 2004b).

Size congruity effect with number words

So far, we discussed the size congruity effect in case of Arabic digits as that is the number representation format with which this effect has been classically and most commonly reported. In the present study, we want to compare congruity effects observed with numerical magnitude and congruity effects observed with scalar adjectives. Having this goal in mind, Arabic digits are not suitable as stimuli since they differ from scalar adjectives not only in their meaning, but also in the fact that digits are presented as one symbol whereas scalar adjectives need to be processed as words before their meaning is accessed. In contrast, number words (i.e., 'three', 'five', etc.) are more comparable to adjectives - they also need to be processed as words before the numerical magnitude is accessed. Therefore, to compare numerical magnitude and scalar adjective meaning processing, we collected data with number words. In this subsection we discuss studies investigating size congruity effect with number words. As discussed below, whether

[5]Note that automaticity of access of numerical (and size) magnitude in general is a prominent line of research of its own (e.g., Dadon & Henik, 2017; Dehaene & Akhavein, 1995; Ford & Reynolds, 2016; Pansky & Algom, 2002; Wong & Szücs, 2013). Here, we are only concerned with automaticity specifically in the sense of processing of the magnitude of the irrelevant dimension in the set-up of the size congruity paradigm.

size congruity effect can be observed with number words still remains an open question, so the present study will add evidence on that question as well.

Most classical models of numerical processing assume that there exists a single representation of analog magnitude codes that can be used for numerical magnitude comparison from symbolic input of various notations; these same magnitude codes would be accessed if the stimuli are presented as e.g., Arabic digits, written number words, spoken number words, etc. (Cipolotti & Butterworth, 1995; Dehaene, 1992; Koechlin et al., 1999; McCloskey, 1992). Nonetheless, empirical evidence shows that there are certain differences in processing different notations that could be attributed to, for example, differences in the amount of experience with a particular notation (Campbell & Epp, 2004), varying processing speed (Cohen Kadosh, Henik, & Rubinsten, 2008) or other factors (see Cohen Kadosh, Henik, & Rubinsten, 2008 for a discussion[6]). Because in the present project, we are interested in processing of number words, we now turn to studies that used size congruity paradigm with number words.

In a numerical comparison task (with size magnitude as the task-irrelevant dimension), a size congruity effect with number words has been reported in English (Foltz, Poltrock, & Potts, 1984), Hebrew (Cohen Kadosh, Henik, & Rubinsten, 2008) as well as with Japanese Kana numbers (syllabic script close to alphabetic script in English, Ito & Hatta, 2003). In contrast, in a physical size comparison task (with numerical magnitude as the task-irrelevant dimension) the results so far are mixed - the congruity effect has not been observed for Japanese Kana numbers (Ito & Hatta, 2003) but has been reported in Hebrew under some conditions (Cohen Kadosh, Henik, & Rubinsten, 2008).[7]

An important aspect that has not been fully taken into account in the previous studies with number words is that the size congruity effect has been shown to be modulated (and masked) by discriminability as well as by variability of the presented stimuli (Algom et al., 1996; Pansky & Algom, 1999). *Discriminability* refers to the psychological difference separating two stimulus values along a dimension, measured in terms of the speed needed to discriminate the two stimuli along this dimension. The second relevant aspect, *variability* refers to the number of different levels of magnitude in each dimension, or how finely grained each dimension is. Both discriminability and variability are thought to influence the salience of each dimension, or the amount of attention that is given to it - the more variable and more discriminable dimension will take more attentional resources. If the irrelevant dimension is more discriminable and variable than the relevant dimension, it will interfere with the relevant dimension simply because it attracted more attentional resources. If the relevant dimension is the more

[6]There has also been a radical counter-proposal - Cohen Kadosh & Walsh, 2009 - which suggested notation-specific representations, but see a wave of counter-arguments in the commentaries published alongside that article.

[7]To our knowledge, there are no published studies that looked at a physical size comparison task with number words in English.

discriminable and variable, the irrelevant dimension will not have an opportunity to interfere because it will not be able to attract enough attentional resources. In their studies, Algom and Pansky demonstrate that only in case discriminability and variability are matched can we conclude that the congruity effect was or was not present specifically due to interference of magnitude codes in each dimension (Algom et al., 1996; Pansky & Algom, 1999). For example, in the study of Ito and Hatta (2003) participants were notably slower in the numerical magnitude comparison task than in the physical size comparison task (the difference was around 250-300 ms), meaning that discriminability was worse for the numerical magnitude than for the physical size in their stimuli. It is then not surprising that they observed a congruity effect when the numerical magnitude was the task-relevant dimension but not when it was the task-irrelevant dimension.[8]

The second study that investigated size congruity effect with the physical size comparison task, by Cohen Kadosh and colleagues (2008), reported the congruity effect both when the numerical magnitude was task-irrelevant and when the size magnitude was task-irrelevant in one of the experiments. In the critical experiment of this study (Experiment 4), the stimuli in two dimensions were matched in terms of variability, but still were not matched in terms of discriminability. In fact, the physical size judgments were faster than the numerical magnitude judgments by around 100-300 ms.[9] Whereas they *do* observe a congruity effect despite this mismatch in discriminability of the two dimensions, the pattern of the effects they observed was somewhat different from that observed for Arabic digits within the same study. Specifically, both congruent and incongruent conditions with number words were in fact slower than a third, neutral condition where the numerical dimension (which was task-irrelevant) did not vary between two stimuli (i.e., same number word presented twice on the screen). In contrast, in the parallel experiment with Arabic digits the neutral condition RT was between the RTs of the congruent and incongruent condition. In addition, their experiment additionally included a numerical distance manipulation for which they observe RT effects in case of digits, but not in case of number words.

Given that in their studies Arabic digits did interfere with size magnitude processing, whereas number words did not interfere with it or did so with a different pattern of effects, Ito and Hatta (2003) as well as Cohen Kadosh and colleagues (2008) propose that Arabic digits and number words differ in their relation to GMS. Either the number words do not have a strong automatic connection to the

[8]Furthermore, they used only two values of physical size (one large and one small), whereas 5 different number pairs were used. This means that the variability of the stimuli was larger for the numerical dimension than for the physical size dimension. However, the stark difference in the discriminability likely made the physical size considerably more salient.

[9]It should be noted that they *did* make the discriminability for physical size dimension more difficult (by making the differences in font size smaller) than in their Experiment 1 to make it more similar to that of numerical magnitude discriminability, but they still did not obtain a full and complete match of discriminability.

GMS representations in this task, or processing number words, unlike Arabic digits, does not recruit GMS representations in general. However, given that neither of these studies fully matched variability and discriminability of the stimuli, more data is needed to make convincing conclusions regarding shared representations of size magnitude and numerical magnitude for number words.

Size congruity effect with conceptual size comparisons

In the present study, we ask whether GMS is recruited for processing magnitude information that is conveyed by scalar adjectives. In this respect, a relevant line of research is the one arguing that GMS is used when comparing the *conceptual size* of objects - e.g., when comparing the (typical) size of a 'lion' and 'ant'.

The size congruity effect has been observed in tasks where participants were presented either with drawings or written names of two objects (e.g., ant lion) and had to choose either the conceptually larger object or the physically larger object (in terms of the size of the presented drawing or the font of the word) with the irrelevant dimension matching or mismatching the relevant dimension (Foltz et al., 1984; Gliksman, Itamar, Leibovich, Melman, & Henik, 2016; Konkle & Oliva, 2012; Paivio, 1975; Rubinsten & Henik, 2002; see also Henik, Gliksman, Kallai, & Leibovich, 2017 for a review). According to Rubinsten and Henik (2002), participants in this task first convert the names of animals into continuous internal representations of their size and subsequently compare these representations in the same way as they would compare numerical magnitude or size magnitude information (see also Gabay et al., 2013, for an argument that the origin of the effect is in the conflict specifically at the level of representations). Thus, they argue that participants were using the same mechanism for comparing conceptual sizes as for comparing magnitudes in perceptual dimensions.

4.1.5 Alternative accounts of the source of the size congruity effect

While the size congruity effect has traditionally been seen as evidence for shared representations underlying numerical and size magnitude (e.g., Cohen Kadosh, Lammertyn, & Izard, 2008; Schwarz & Heinze, 1998), several alternative accounts of the observed effects have been brought up. In order to conclude that the representations are shared between the two dimensions in our own set of experiments, we have to address these alternative explanations.

One alternative account of the size congruity effect is that participants assign verbal labels to the stimuli that they see on the screen — e.g., labels 'large' and 'small' are assigned both to digits and physical sizes. The conflict then arises between the verbal labels (e.g., 'large' for digit and 'small' for physical size in the incongruent condition). If this were the case, we could not conclude that representations of numerical magnitude and physical size are shared based on the

size congruity effect. However, there are several reasons to exclude this possibility. First, if the size congruity effect were purely due to assigning conflicting verbal labels to the stimuli, this conflict should result in the same congruity effect regardless of how large the differences between magnitudes in the irrelevant dimension are. In other words, we would not expect the congruity effect to be modulated by the specific magnitude differences in the irrelevant dimension. However, as we have discussed, we know that larger differences in the task-irrelevant dimension in fact result in a larger congruity effect (i.e., more interference; e.g., Cohen Kadosh, Henik, & Rubinsten, 2008; Henik & Tzelgov, 1982; Santens & Verguts, 2011). Second, the size congruity effect has also been observed in a similar subliminal priming paradigm where no verbal labeling was possible (Lourenco et al., 2016). Therefore, we conclude that this is an unlikely explanation of the size congruity effect.

Another alternative account is based on visual attention capture (Risko et al., 2013). This account is based on the observation from vision research that larger items in a scene capture attention more than small items. Given that in a size congruity paradigm one item is visually larger than the other, this item will capture attention first and, therefore, might have a temporal advantage - it could be processed first. Thus, if the task is to react to the item with the larger numerical magnitude, in the congruent condition the physically larger item is at the same time the target for the numerical magnitude task, and thus target item identification would be boosted (relative to the incongruent condition). However, this account can only explain the size congruity effect in the numerical comparison task (with size magnitude as the task-irrelevant dimension), whereas the size congruity effect is also observed in the physical comparison task (with numerical magnitude as the task-irrelevant dimension). Specifically, if the congruity effect was explained solely by temporal processing advantage of the larger object on the screen, we would not expect to observe a congruity effect when the participants' task is to indicate the larger physical size object because then the larger object would be processed first in both conditions. Moreover, as noted by Arend and Henik (2015) this account is meant to explain the size congruity effect when the participants are instructed to choose the numerically larger item, whereas if participants are instructed to choose the numerically smaller item, it predicts a reverse effect - faster responses in the incongruent condition (since attention would still be captured by the larger item first). However, in contrast to this prediction, the size congruity effect is observed with the 'choose smaller' instructions too albeit somewhat smaller in effect size (Arend & Henik, 2015; Tzelgov et al., 1992). Arend and Henik (2015) argue that, given the reduction in congruity effect for 'choose smaller' instructions, there does seem to be some effect of attention capture in the size congruity task, but it clearly cannot fully explain the congruity effect (a position also accepted as a possibility by Risko et al., 2013 who proposed the account based on visual attention capture). Hence, the attention capturing cannot be taken as the full explanation of the size congruity effect.

The most important and relevant alternative account suggests that the size congruity effect originates in the decision (i.e., response selection) stage of processing (Faulkenberry, Cruise, Lavro, & Shaki, 2016; Santens & Verguts, 2011; see also Proctor & Cho, 2006 for another account with similar reasoning). This account is based on the simple fact that in the congruent condition both the task-relevant and the task-irrelevant dimensions (size and numerical magnitude) converge on the same (potential) (motor) response (e.g., *right larger* or *left larger*), whereas in the incongruent condition the relevant and irrelevant dimensions diverge on different (motor) responses. One can imagine that processing of numerical magnitude and size magnitude happens in parallel, using different representations, but both result in a potential motor response option. These motor responses then compete to for selection. Importantly, a computational implementation of this account (Verguts et al., 2005) also suggests an explanation for the previously mentioned modulation of the congruity effect by the difference between magnitudes in the task-irrelevant dimension. According to this model, the amount of activation passed on to the units deciding between alternative motor responses (decision units) depends on the difference between magnitude values from which the system was choosing. When the difference between them is large, there will be a stronger activation passed on to the potential motor response and this activation will thus have a stronger influence on the decision unit. As a result, when the difference in the task-irrelevant dimension is large, there will be a stronger activation of the response induced by this dimension on the decision units than when the difference on the task-irrelevant dimension is small. Thus, the larger difference on the task-irrelevant dimension will have a stronger impact on the decision units, delaying the decision for the eventual response in the task relevant dimension, and causing a larger congruity effect (see Verguts et al., 2005, for details).

There are several counter-arguments against an account that is exclusively based on the conflict at the decision stage of processing (henceforth, referred to as 'decision stage conflict'). First, such an account of the congruity effect (as presented by Santens & Verguts, 2011) predicts that it should arise to an equal extent with different decision polarities (i.e., 'choose smaller' task or 'choose larger' task) and with different task-relevant and task-irrelevant dimensions, as long as in each case there are two response options compatible with both task-relevant and task-irrelevant dimensions. However, the size congruity effect seems to be modulated by the decision polarity ('choose larger' or 'choose smaller', as already mentioned above) and differs depending on which dimension is task-relevant (i.e., numerical comparison or physical size comparison task; Arend & Henik, 2015; Tzelgov et al., 1992; see Arend & Henik, 2015 for this argument and supporting evidence). Moreover, ERP studies on the size congruity effect found that a neural correlate of interference is observable both at an early stage of processing (150-250 ms after stimulus presentation), the point when the stimuli are thought to be mapped to magnitude representations, and later stage of processing (300-430 ms), the point

when the response is thought to be selected (Szucs & Soltesz, 2008; see also Cohen Kadosh et al., 2007; Schwarz & Heinze, 1998 for converging evidence). While it is difficult to pinpoint the source of an ERP effect, these findings provide evidence that at least part of the congruity effect arises from a conflict at an early processing stage, possibly at level of magnitude representations.

Note that it is also possible that the size congruity effect arises partially due to a conflict at the decision stage of processing and partially due to a conflict at shared representations of size magnitude and numerical magnitude (this has also been suggested by proponents of the response selection account - e.g., Faulkenberry et al., 2016; Santens & Verguts, 2011). In the present study, we collect additional data with the same stimuli but a completely different task to be able to test whether the observed congruity effect originates exclusively from the conflict at the decision stage of processing.

4.1.6 Present study

In the present series of experiments, in a first step we use the size congruity paradigm to look at the congruity effect between numerical magnitude conveyed by number words and the physical (font) size magnitude of these number words. One group of participants performed a numerical magnitude comparison (Experiment 1a); another group of participants performed a physical size comparison task (i.e., font size comparison; Experiment 1b) on the same stimuli. As discussed above, the existing studies investigating size congruity effect with number words had unbalanced stimuli in terms of variability and discriminability of magnitudes in the task-relevant and task-irrelevant dimensions. In the present experiments, we balanced variability and discriminability of the stimuli, making it a stronger test case for potential congruity effects than existing studies with number words. We collected data with number words (and not digits) to be able to compare the observed effects with those for scalar adjectives that we are primarily interested in in the present study.

In the next step, we use the reasoning and the experimental set-up of the size congruity paradigm to look at a potential representational overlap between the meaning of scalar adjectives and magnitude representations in GMS. We did so by inspecting the potential interference between the retrieval of the (meaning of) scalar adjectives and presented physical size magnitude. These experiments were parallel to the ones with number words. One group of participants performed a comparison of pairs of scalar adjectives (e.g.,'kort-lang' ['short-long'], 'laag-hoog' ['low-high'], 'licht-zwaar' ['light-heavy']) in terms of their meaning (Experiment 2a). Specifically, they were asked to judge which of two antonymous adjectives "means more/less of something" while the match with the task-irrelevant font size of these adjectives was manipulated. Henceforth, we refer to the scalar adjective comparison (Experiment 2a) and numerical magnitude comparison (Experiment 1a) as *semantic comparison* tasks. Another group of participants performed a

physical size comparison with pairs of scalar adjectives as stimuli (Experiment 2b). Again, the match with the meaning of scalar adjectives was manipulated to create congruent and incongruent trials. Collecting data for number words and scalar adjectives in experiments with parallel designs allows us to compare these two symbolic references to magnitudes. If scalar adjectives and number words make use of GMS representations in the same way, we expect to see parallel congruity effects for both. Alternatively, they may differ either in automaticity or in the source of congruity effect.

To anticipate, we find a reliable congruity effect in case of the semantic comparison tasks with both number words and scalar adjectives (i.e., with the size magnitude being the task-irrelevant dimension), though not in case of physical size comparison tasks. In order to locate the source of this congruity effect (representational overlap vs. decision stage conflict), we followed up these experiments with two additional experiments (Experiment 1c for number words and Experiment 2c for scalar adjectives). These experiments used a different task which asked participants to indicate whether the two presented number words or scalar adjectives were same (e.g., 'one-one') or different (e.g., 'one-six'), i.e., they performed a 'same'/'different' judgment. The stimuli in the 'different' trials (i.e., trials with two different number words or scalar adjectives) were the same pairs as the ones used in the comparison experiments (Experiments 1a, 1b, 2a, 2b). These were the trials of interest that we analyzed.

For the same/different task, the shared representations and the decision stage conflict accounts make different predictions for the critical 'different' trials. Specifically, because in this task the response options are 'same' and 'different', under the decision stage conflict account the size magnitude dimension and the numerical magnitude (or adjective meaning polarity) dimension would compete for one of these response options. Given that in the 'different' trials two different number words (or adjectives) along with two different physical sizes were presented in both congruent and incongruent trials (i.e., trials considered 'congruent' and 'incongruent' in the comparison tasks), both dimensions should activate the 'different' response in both type of trials. So no conflict should arise between potential responses from the two dimensions in either type of the trials. Thus, the decision stage conflict account predicts that no congruity effect should be observed in the same/different task. In contrast, because the shared representations account claims that the congruity effect arises from the magnitude code mapping stage of the processing, it still predicts a congruity effect in this task - processing mismatching numerical magnitude and size magnitude should result in a conflict at the level of representations regardless of for which exact task (goal) the participant is computing these representations. Thus, according to the shared representations account we should still observe faster reaction times in trials congruent than in incongruent trials.

To describe the reasoning presented above in a different way, let us consider what happens in the same/different task in case of trials congruent in terms of

magnitude and trials incongruent in terms of magnitude. Let us take an example of a trial congruent in terms of magnitude - e.g., '**five** three', - and an example of a trial incongruent in terms of magnitude - 'five **three**'. According to the decision stage conflict account, the RTs are longer when two dimensions (numerical magnitude and size magnitude dimensions) activate different response options than when they activate the same response option. In this case, in both types of trials both dimensions would activate a 'different' response. Thus, there will be no competition for activation of the response in the trials incongruent in terms of magnitude. Accordingly, the decision stage conflict account predicts that the RTs in the trials incongruent in terms of magnitude should be identical to the RTs in the trials congruent in terms of magnitude. Recall that according to the representational overlap account, there is a single shared magnitude representation code that the numerical magnitude and size magnitude compete for/activate. Let us consider what happens in the trial incongruent in terms of magnitude according to this account. If we, for the sake of the example, assume that the shared magnitude code has left-right directionality, then 'five' in our example will claim a position to the right of '**three**' in the numerical dimension, but to the left of '**three**' in the physical size magnitude. This 'competition for positions' on the underlying analog dimension should prolong RTs in the trials incongruent in terms of magnitude relative to the trials congruent in terms of magnitude (where there will be no such 'competition for positions') in absence of any response competition. Thus, under the representational overlap account we expect to observe shorter RTs in the trials congruent in terms of magnitude than in the trials incongruent in terms of magnitude. The same reasoning applies with scalar adjectives as stimuli.

4.2 Experiments 1a and 1b: comparison tasks with number words

Participants saw pairs of number words on the screen and were asked to decide which is numerically larger/smaller (Experiment 1a) or which is presented in larger/smaller font size (Experiment 1b). These experiments follow the classical size congruity paradigm (Besner & Coltheart, 1979; Henik & Tzelgov, 1982) except that instead of digits number words were presented. The same stimuli were presented in both experiments. The number word with a larger numerical magnitude could be presented in a large font size, creating a *congruent* condition, or in a small font size, creating an *incongruent* condition (and correspondingly with the smaller numerical magnitude). We expected to observe a congruity effect - shorter reaction times in the congruent condition than in the incongruent condition. Such a congruity effect would suggest that the magnitude of the task-irrelevant dimension was automatically processed and that it interfered with processing of the magnitude of the task-relevant dimension.

The participants either indicated which of the two items of a trial is larger or they indicated which is smaller in the task-relevant dimension, a manipulation that we will refer to as *decision polarity*. The decision polarity was reversed for every participant in the middle of the experiment.

The data for all experiments reported in this manuscript have been collected remotely — participants completed the experiments from their own computers in a web browser. Previous studies testing the difference between reaction times observed from an experiment running in a web browser and using traditional lab tools (such as Matlab Psychophysics Toolbox) showed that although there was a time-lag in the reaction times observed in a study running in a web browser (of about 25 ms), there was no difference in terms of the distributions of the RTs and no difference in sensitivity to RT-differences between experimental conditions (de Leeuw & Motz, 2016; Reimers & Stewart, 2015). When it comes to within-participant designs, potential effects should be detected with the same reliability as with traditional lab tools because the equipment stays the same throughout the experiment. A number of classical effects in cognitive psychology have been successfully replicated with data collected online (e.g., Crump et al., 2013; Semmelmann & Weigelt, 2017; Zwaan, Pecher, et al., 2018), leading to the conclusion that online data collection is a suitable and reliable option for hypothesis testing. Finally and most importantly, a recent study which used specifically the size congruity paradigm in web browsers observed data quality comparable to the physical lab-based studies and successfully replicated the classical congruity effects (Kochari, 2019).

4.2.1 Method

Participants

Participants for these and all other experiments reported in this manuscript were recruited via Prolific.ac (Palan & Schitter, 2018). All our experiments were in Dutch. To take part in the experiments, participants had to be 18-35 years old, native speakers of Dutch, and born in and currently living in the Netherlands. Each participant was told that the study will take approximately 20 minutes and was reimbursed for their time with 3.50 British pounds. After data collection, the following participant exclusion criteria were applied: a participant gave incorrect responses in more than 15% of trials, the time spent reading the first instructions of the experiment was less than 10 seconds, the time spent on the whole experiment was longer than 30 minutes (measured from when they started the first practice trial). These criteria were applied to ensure that the participants included in the analysis definitely understood the instructions and did not devote time to another task (e.g., opening another website) during the experiment. For each experiment, data collection continued until we reached the desired number of participants meeting the inclusion criteria.

We aimed to collect data from the same number of participants across the experiments with number words and scalar adjectives. The size congruity effect is typically robust and detectable with relatively few participants: previous studies report significant effects with 10-20 participants (e.g., Cohen Kadosh, Henik, & Rubinsten, 2008; Henik & Tzelgov, 1982; Kaufmann et al., 2005; Santens & Verguts, 2011). We do not know if the effect size in case of the scalar adjectives will be comparable to the one for the numerical magnitudes, it may in fact be smaller than for numerical magnitude. Given these considerations, we decided to collect data from 50 participants in each of the experiments.

Fifty-five participants completed Experiment 1a, i.e., the semantic comparison task with number words. Four participants were excluded from the analysis because they read the first instructions in less than 10 s. One further participant was excluded because they spent more than 30 minutes on the experiment. The mean age of the included participants was 25 years (SD 4.6; 31 male and 19 female). On average, they took approximately 14:40 minutes to complete the experiment (SD 02:22, min. 12, max. 26).

Fifty-eight participants completed Experiment 1b, i.e., the physical size comparison task with number words. Five participants were excluded from the analysis because they gave incorrect responses in more than 15% of trials. Two participants were excluded because they spent more than 30 minutes on the experiment. Finally, one participant was excluded because they read the first instructions in less than 10 s. The mean age of the included participants was 25 years (SD 4.6; 32 male and 18 female). On average, they took approximately 14:50 minutes to complete the experiment (SD 02:37, min. 12, max. 23).

Stimuli

Exactly the same stimuli were used across the experiments on semantic comparison (1a) and physical size comparison (1b) with the only difference between them being in the instructions participants received (see *Procedure* below for details).

We used five pairs of number words: 'een-zes' ['one-six'], 'twee-acht' ['two-eight'], 'twee-vijf' ['two-five'], 'drie-acht' ['three-eight'], and 'vier-acht' ['four-eight'], presented in five combinations of font sizes respectively: 41-47 pt, 37-42 pt, 41-46 pt, 38-42 pt, 43-48 pt. In other words, for example, in case of the pair 'een-zes', 'een' was presented in font size 41 pt and 'zes' was presented in font size 47 pt in the congruent condition and vice versa in the incongruent condition. Each number word pair was matched with a unique font size pair in order to ensure equal variability in both dimensions. Both number words within a pair had the same number of letters in order to avoid a potential confound with the visual difference in the length of words.

Similarly to previous studies (e.g., Algom et al., 1996; Pansky & Algom, 1999; Santens & Verguts, 2011), comparable discriminability in the task-relevant and task-irrelevant dimensions was achieved by matching the mean reaction time ob-

served for comparison of the number words when both number words of a given pair were presented in the same font size (in our case, this was 44 pt) and for comparison of the font sizes in a meaningless context (in our case, strings of consonants were presented in different font sizes). We collected data in a norming study prior to the experiments from 30 participants recruited from the same population (none of these participants subsequently took part in the actual experiments). For details on this norming study, see supplemental online materials. The mean RTs and error rates for the selected number word and font size combinations are provided in *Table 4.1.*[10]

Table 4.1: Mean RT (SD) and error rate observed in stimuli pre-test for the selected pairs of number word and font size combinations.

number word pair	**RT (SD)**	**error rate**	**font size pair**	**RT (SD)**	**error rate**
'een-zes'	701 (135) ms	0%	41-47 pt	720 (222) ms	2.43%
'twee-acht'	743 (157) ms	1%	37-42 pt	747 (229) ms	1.74%
'twee-vijf'	780 (202) ms	0.67%	41-46 pt	774 (222) ms	5.56%
'drie-acht'	790 (170) ms	3.67%	38-42 pt	761 (254) ms	5.92%
'vier-acht'	810 (180) ms	1.67%	43-48 pt	787 (235) ms	6.97%
across all pairs	764 (174) ms	1.4%	across all pairs	757 (233) ms	4.52%

The five stimulus pairs of interest were intermixed with three filler stimulus pairs in order to reduce the possibility that participants will learn responses to specific pairs that they observed. These filler pairs were 'twee-drie' ['two-three'], 'zeven-negen' ['seven-nine'], and 'drie-vier' ['three-four'] presented in font sizes 42-46 pt, 38-43 pt, 38-44 pt respectively. In case of filler trials, the discriminability was not matched.

Each of the number word pairs was presented in the congruent (numerically larger number word presented in larger font size) and the incongruent (numerically larger number word presented in smaller font size) condition an equal number of times. Examples of displays in the congruent and incongruent conditions are provided in *Figure 4.1.* Each number word in a pair appeared on both sides of the screen in each condition. Each configuration (of congruity and location on the screen) was repeated five times. Finally, participants performed a 'choose larger'

[10]In the norming study, we collected data for 12 different number word pairs (total number of possible number word pairs in Dutch with equal length of words within a pair) as well as 21 different font size combinations. However, we were able to match RTs of physical size comparison and number word comparison only in case of five items. That is because the general speed of comparison was substantially faster for physical size comparisons than for numerical magnitude comparisons. See the supplemental online materials for RTs and error rates for all number word pairs and font sizes that we tested. Note that in order to achieve longer RTs in the physical size comparisons we would have to make the difference in the fonts smaller than the ones we already tested, but it was not possible since smaller differences in font sizes are at some point not clearly visible anymore and result in unacceptably high error rates. We compensate for the low number of different stimuli pairs by administering a large number of trials per participant.

Figure 4.1: Examples of displays in congruent and incongruent conditions in Experiments 1a and 1b.

(a) congruent condition (b) incongruent condition

as well as 'choose smaller' tasks (decision polarities). In total, thus, participants saw 8 (number word pairs; 5 pairs of interest and 3 filler pairs) * 2 (levels of congruity) * 2 (sides of the screen) * 5 (repetitions) * 2 (decision polarities) = 320 trials. Out of these trials, 200 were trials of interest and 120 were filler trials. Out of trials of interest, 100 trials were on 'choose larger' and the other 100 on 'choose smaller' decision polarity. Within each of the decision polarities, participants saw 50 trials of interest in the congruent condition and 50 trials of interest in the incongruent condition. In each experiment, half of the participants performed the 'choose larger' task first (160 trials after which they were instructed to make decisions with the other decision polarity) and half of the participants performed the 'choose smaller' task first.

Procedure

The experiments were administered using jsPsych, a JavaScript library for running behavioral experiments in a web browser (https://www.jspsych.org/; de Leeuw, 2015).

In Experiment 1a, participants were instructed to indicate the side of the screen with a larger or smaller number (i.e., numerical magnitude) by pressing a corresponding key on their keyboard. They were told to ignore any other properties of the display. An example was given to demonstrate that it is indeed the numerical magnitude that they should pay attention to. In Experiment 1b, participants were instructed to indicate the side of the screen with a word in larger or smaller font size. In this case too, there was an example showing that they should ignore the numerical magnitude and only pay attention to the font size. Participants were asked to keep their index fingers on two response keys 'P' and 'Q' and encouraged to respond as soon as possible.

Participants opened the page with the experiment by clicking on a link on the Prolific.ac website. They first read the consent form information and agreed to participate. They were then presented with instructions for the first decision polarity. At this point, the participants were not informed that they will later

be asked to make a decision with the reversed polarity. After reading the first instructions, they had a chance to practice the experimental task in four practice trials with stimuli which did not appear in the actual experiments. During the practice trials, they received feedback on whether the given response was correct. The experimental trials of the first decision polarity then followed. There was no feedback given at this stage. The experimental trials were presented in a random order without restrictions, divided into two blocks. There was a break between the blocks. Next, the participants were informed that in the second half of the experiment they will be performing a judgment with the opposite polarity, using the same keyboard keys. They again had a chance to practice, this time on seven practice trials. In the second half of the experiment they again saw trials in a random order without restrictions, divided into two blocks.

Each experimental trial started with a fixation cross in the center of the screen displayed for 200 ms. It was then replaced by the two stimuli displayed to the left and the right of the middle of the screen for 2000 ms or until the participant pressed a response button. The response was given by pressing either 'P' on the keyboard if the stimulus on the right side was the intended response or 'Q' if the stimulus on the left side was the intended response. If no response was given within 2000 ms, the trial ended automatically. The interval between the response and onset of the fixation cross of the next trial was 200 ms. In order to reduce effects of anticipating the upcoming stimulus (e.g., Clementz, Barber, & Dzau, 2002), the interval between the display of the fixation cross and the display of the two trial stimuli was varied randomly between trials - each time it was a random number between 300 and 700 ms. For the same reason, we also added filler trials that were empty (the fixation cross was followed by a blank screen for 500 ms and no response was required) to the experiment. We added 12% of empty filler trials to each experiment, and participants were informed about the presence of such trials in the instructions.

All stimuli were presented in *Courier* monospace font. The distance between the point where the word on the left ended and the center of the screen was equal to the distance between the center of the screen and the point where the word on the right started. This distance was same for all trials.

Analysis

Only trials in which a correct response was given were included in the analysis of the reaction times. In addition, we excluded all trials in which the RT was shorter than 200 ms as those were likely accidental button presses. Finally, we also excluded all trials in which the RT was longer than the mean RT plus three standard deviations in a given decision polarity for a given participant.

The analysis described here was also used for all other experiments in the present study, so we describe it in detail. Data were analyzed in the R environ-

ment (R Core Team, 2020)[11] and inferences were made by fitting linear mixed effect models using functions in the package *lme4* (Bates et al., 2015). The LME models always included fixed effects of the factors *congruity* (congruent vs. incongruent), *decision polarity* ('choose larger' vs. 'choose smaller') and their interaction. Initially, we fitted a model with a maximal random effect structure (Barr, Levy, Scheepers, & Tily, 2013), i.e., allowing for by-participant and by-item (i.e., number word pair or adjective pair in further experiments) intercepts as well as varying slopes for each effect. Whenever the maximal model did not converge or resulted in a singular fit, we gradually simplified the random effect structure of the original maximal model by excluding the random effect that accounted for least variance until a non-singular converging model was reached (following one of the recommendations of Barr et al., 2013). The reported p-value for each factor was obtained using the Satterthwaite approximation for denominator degrees of freedom as implemented in the R package *lmerTest* (Kuznetsova et al., 2017).

In addition to the frequentist LME models, we also fit parallel Bayesian multilevel models using the package *brms* (Bürkner, 2017, 2018). These models allowed us to quantify how much our data supports the null or the alternative hypothesis (see Nalborczyk, Batailler, Lœ venbruck, Vilain, & Bürkner, 2019; Nicenboim & Vasishth, 2016; Vasishth, Nicenboim, Beckman, Li, & Kong, 2018 for descriptions of Bayesian multilevel models in the context of psycholinguistic research). We chose an ex-gaussian distribution model because it provides a considerably better fit for reaction time data which is typically (and also clearly in the present studies) right-skewed (Lindelø v, 2020; Rousselet & Wilcox, 2019). In addition, examination of posterior predictive values generated by models with a gaussian distribution and an ex-gaussian distribution showed that the latter model was overwhelmingly better able to predict values close to the data we observed. The random effects structure was maximal as described above. We used a normally distributed prior with mean 0 and standard deviation 100 ms for the population-level (i.e., fixed) effects. Such a prior meant that we were 95% certain that the effect of congruity, task and interaction was between -200 and 200 ms.[12] The priors for the remaining parameters were left as default. The models were fit with four chains and 5000 iterations half of which were the warm-up phase. Model convergence was verified by making sure that there were no divergent transitions, Rhat values were close to one, and by examining the trace plots.[13] We inspected

[11]Specifically R version 3.6.3 was used along with the following packages: *brms* (version 2.12.0; Bürkner, 2017, 2018); *ggplot2* (version 3.3.0; Wickham, 2016); *Hmisc* (version 4.4-0; Harrell Jr, 2020); *knitr* (version 1.28; Xie, 2014); *lme4* (version 1.1-21; Bates, Mächler, Bolker, & Walker, 2015); *lmerTest* (version 3.1-1; Kuznetsova, Brockhoff, & Christensen, 2017); *Matrix* (version 1.2-18; Bates & Maechler, 2019); *plyr* (version 1.8.6; Wickham, 2011); *Rcpp* (version 1.0.4; Eddelbuettel & Balamuta, 2018) *readr* (version 1.3.1; Wickham, Hester, & Francois, 2018);

[12]We ran additional models with population-level effect prior SDs 200 and 400. Because the estimates resulting with these priors were extremely close to those with SD 100, we do not report them here. Full results of these models are available in supplemental online materials.

[13]An additional recommended criterion of convergence is effective sample sizes of at least

mean estimates for the effects of interest along with 95% credible intervals (CrI) of the posterior estimate. The 95% CrI should be interpreted as containing the true value of the effect with 95% probability. To quantify the evidence provided by the data for or against the effects of interest being zero, we calculated Bayes Factor values using Savage–Dickey density ratio method (Dickey & Lientz, 1970; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). Our null hypothesis here was that the effect is exactly zero, whereas the alternative hypothesis was that the effect is not exactly zero (note that this is a two-sided test). This calculation gave us a BF_{01} (how much the collected data increases our confidence that the effect is exactly zero relative to how confident we were about it before the data was collected, i.e., relative to the prior described above) or vice versa, BF_{10} (how much the collected data increases our confidence that that the effect is *not* exactly zero relative to how confident we were about it before the data was collected). For example, BF_{01}=1 means that the collected data does not change our confidence about the effect being zero (i.e., zero was equally likely in the prior and posterior distributions), BF_{01}=2 means that the collected data should double our confidence in that the effect is zero, and BF_{10}=2 means that we should double our confidence in that the effect is not zero. Note that we report BF_{01} or BF_{10} depending on which one was larger. We interpreted BFs below 3 as inconclusive, above 3 as moderate evidence and BFs above 10 as strong evidence in favor of one hypothesis over another (Jeffreys, 1998).

Raw data, analysis scripts and full model results for all experiments presented in this manuscript are provided in the supplemental online materials available on Open Science Framework under `https://osf.io/kh6eb/`.

4.2.2 Results

In Experiment 1a, i.e., the semantic comparison task with number words, participants included in the analysis made 3.58% errors in the whole experiment on average (min. 0%, max 10%). Data cleaning procedure resulted in exclusion of RTs of 2.92% of trials of interest (excluded incorrect responses are also counted here). The resulting mean RTs and error rates per congruity overall and in each decision polarity are given in *Table 4.2* and visually depicted in *Figure 4.2a*. Mean RTs and error rates per number word pair across the decision polarities are given in *Table 4.3*.

The linear mixed effect model with maximal random effect structure for Experiment 1a data resulted in a singular fit. The random effect structure was gradually

10% of the total number of post warm-up samples (Vasishth, Nicenboim, et al., 2018). This was not the case for one of the parameters in most of the models that we fit. Specifically, for the correlation between congruity and decision polarity slopes by participant the lowest number was 6%. However, the Rhat values were 1.01 and the effective sample size increased linearly with increasing number of iterations. Therefore, we concluded that the mixing was sufficient (*Brief Guide to Stan's Warnings*, 2020).

simplified to achieve a converging non-singular model fit. The final model included varying intercepts per-participant and per-item (i.e., per number word pair) as well as varying slopes for the effect of decision polarity in both cases. There was a significant main effect of congruity ($\beta = 29, SE = 4.0, t = 7.37, p < 0.0001$) and a significant main effect of decision polarity ($\beta = 43, SE = 16.7, t = 2.6, p = 0.038$). The interaction effect was not significant ($\beta = -1, SE = 5.7, t = -0.22, p = 0.82$). For this and all further analyses, the result of the maximal random effect structure model (resulting in a singular fit) did not contradict the results of the model with the simplified random effect structure; results of all models can be inspected in the supplemental online materials.

The Bayesian LME model estimated for the main effect of congruity $\hat{\beta} = 25$ ms, 95% CrI = [15.93 35.64], BF_{10}=14.3; for the main effect of decision polarity $\hat{\beta} = 27$ ms, 95% CrI = [-14.95 69.09], BF_{01}=2.9, for the interaction between congruity and decision polarity $\hat{\beta} = 3$ ms, 95% CrI = [-11.52 17.93], BF_{01}=29.2. Thus, there was strong evidence that the congruity effect was not zero, no clear evidence for or against the decision polarity effect being zero (though most of the weight of the posterior distribution is on one side of zero, so for the alternative hypothesis) and strong evidence that the interaction between congruity and decision polarity was zero.

Table 4.2: Mean RT (SD), error rate overall and for each decision polarity in Experiment 1a, semantic comparison with number words.

decision polarity	**congruent**	**incongruent**
overall	720 (182) ms, 1.88%	750 (184) ms, 2.16%
'choose larger'	699 (166) ms, 1.96%	729 (172) ms, 2.56%
'choose smaller'	742 (195) ms, 1.80%	772 (194) ms, 1.76%

Table 4.3: Mean RT (SD), error rate per number word pair (both decision polarities) in Experiment 1a, semantic comparison with number words.

number word pair	**congruent**	**incongruent**
'een-zes'	646 (137) ms, 0.2%	672 (152) ms, 0.3%
'twee-acht'	697 (176) ms, 0.7%	734 (166) ms, 1.3%
'twee-vijf'	708 (168) ms, 0.8%	735 (174) ms, 0.7%
'drie-acht'	768 (193) ms, 2.3%	799 (189) ms, 3.8%
'vier-acht'	788 (197) ms, 5.4%	817 (203) ms, 4.7%

In Experiment 1b, i.e., the physical size comparison task with number words, participants included in the analysis made 5.78% errors in the whole experiment on average (min. 2%, max. 12%). Data cleaning procedure resulted in exclusion

of RTs of 8.36% of trials of interest (excluded incorrect responses are also counted here). The resulting mean RTs and error rates per congruity overall and in each decision polarity are given in *Table 4.4* and visually depicted in *Figure 4.2b*. Mean RTs and error rates per number word pair across the decision polarities are given in *Table 4.5*.

The LME model with maximal random effect structure for Experiment 1b also resulted in a singular fit. The random effect structure was gradually simplified to achieve a converging non-singular model fit. The final model included a per-participant intercept allowing for varying slopes for the effect of decision polarity and allowed for varying random slopes for the congruity effect by-item. None of the effects were significant (main effect of congruity - $\beta = 44, SE = 67.4, t = 0.65, p = 0.54$; main effect of decision polarity - $\beta = 22, SE = 15.7, t = 1.44, p = 0.15$; interaction of congruity and decision polarity - $\beta = 7, SE = 8.4, t = 0.86, p = 0.38$). Note that whereas in the overall means there does seem to be a difference in RTs between congruent and incongruent conditions, closer inspection of the RTs observed for each of the number word pairs (as can be seen in *Table 4.5*) shows that in case of two number word pairs the RTs were in fact shorter for the incongruent than for the congruent condition. This is reflected in the non-significant congruity effect in the model.

The Bayesian LME model estimated for the main effect of congruity $\hat{\beta} = 20$ ms, 95% CrI = [-57.28 94.46], BF_{01}=2.2; for the main effect of decision polarity $\hat{\beta} = 10$ ms, 95% CrI = [-36.26 53.72], BF_{01}=4.32, for the interaction between congruity and decision polarity $\hat{\beta} = 6$ ms, 95% CrI = [-31.23 42.98], BF_{01}=6.06. Thus, there is no clear evidence for or against the congruity effect being zero, moderate evidence that decision polarity effect is zero and moderate evidence that the interaction between congruity and decision polarity is zero.

Table 4.4: Mean RT (SD), error rate overall and for each decision polarity in Experiment 1b, physical size comparison with number words.

decision polarity	**congruent**	**incongruent**
overall	744 (241) ms, 5.06%	788 (259) ms, 8.62%
'choose larger'	733 (239) ms, 5.12%	775 (259) ms, 7.64%
'choose smaller'	757 (243) ms, 5.00%	803 (260) ms, 9.60%

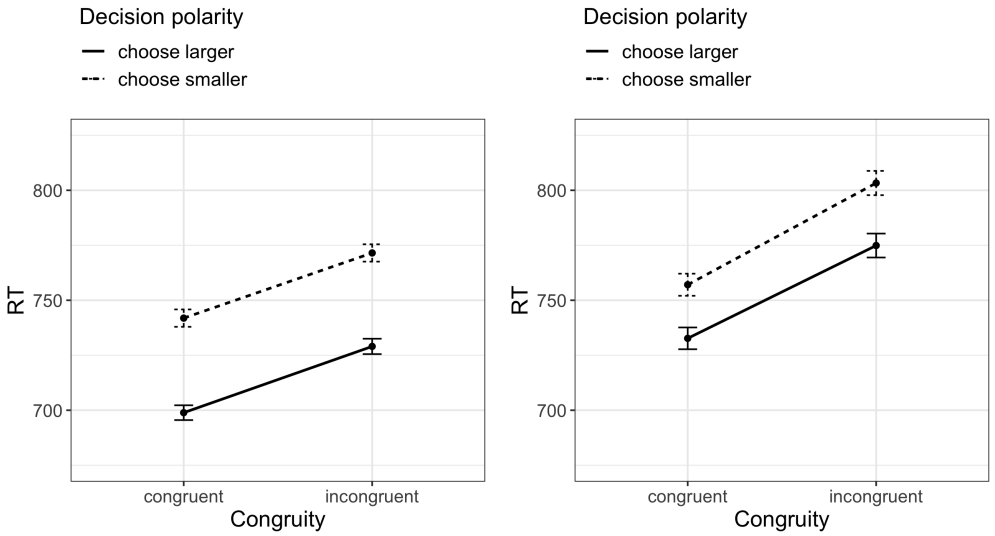
4.2.3 Interim discussion

Let us first consider the implications of the results of the numerical magnitude comparison task (Experiment 1a). In this task, we observed a clear congruity effect that was stable across different number word pairs. Observing the congruity

Table 4.5: Mean RT (SD), error rate per number word pair (both decision polarities) in Experiment 1b, physical size comparison with number words.

number word pair	**congruent**	**incongruent**
'een-zes'	690 (198) ms, 1.60%	797 (246) ms, 7.51%
'twee-acht'	721 (211) ms, 2.10%	748 (235) ms, 2.80%
'twee-vijf'	663 (186) ms, 1.10%	936 (314) ms, 26.60%
'drie-acht'	853 (290) ms, 15.2%	751 (238) ms, 2.70%
'vier-acht'	821 (262) ms, 5.30%	756 (235) ms, 3.50%

Figure 4.2: Mean RTs per congruity and decision polarity in comparison tasks with number words. The error bars depict the standard error value.



(a) Experiment 1a: semantic comparison task with number words.

(b) Experiment 1b: physical size comparison task with number words.

effect here is consistent with previous studies that administered the size congruity paradigm with number words (Cohen Kadosh, Henik, & Rubinsten, 2008; Ito & Hatta, 2003). Importantly, unlike the previous studies, we have matched the stimuli in the task-relevant and task-irrelevant dimensions in terms of both variability and discriminability. In addition, we collected our data in a language for which the size congruity effect with number words has not previously been reported - Dutch. Thus, these results support the robustness of the size congruity effect in the numerical comparison task with number words.

The reaction times were descriptively shorter in trials where the participants were asked to choose a numerically larger number word than in the trials where the participants were asked to choose a numerically smaller number word; this effect was significant in the frequentist LME but inconclusive in the Bayesian LME

estimates. In general, the shorter RTs for the 'choose larger' decision polarity is consistent with the pattern previously reported for Arabic digits (Arend & Henik, 2015). Importantly, we observed a congruity effect for both decision polarities. This is consistent with previous studies that administered both decision polarities in the size congruity task with Arabic digits (Arend & Henik, 2015; Tzelgov et al., 1992). To our knowledge, no previous studies administered different decision polarities with number words, so this is a first demonstration of the congruity effect with the 'choose smaller' decision polarity.

However, unlike in this previous study with Arabic digits (Arend & Henik, 2015; Tzelgov et al., 1992), the size of the congruity effect in our experiment was not modulated by the polarity of instructions. In the studies with Arabic digits, a larger congruity effect was reported for the 'choose larger' decision polarity than for the 'choose smaller' decision polarity. We, on the contrary, have strong evidence that the interaction of congruity and decision polarity is zero in our data.

Let us now turn to the results of the physical size comparison task (Experiment 1b). In this task, the difference between the congruent and incongruent conditions was not consistent across different number word pairs. We have inconclusive evidence for or against the congruity effect being zero (BF_{01}=2.2) though the null hypothesis is supported by the data slightly more than the alternative hypothesis. The lack of a significant congruity effect in this task is consistent with results of some previous studies (Ito & Hatta, 2003; Cohen Kadosh, Henik, & Rubinsten, 2008, Experiment 1). Recall that there was only one study to date reporting a size congruity effect with number words in the physical size comparison task (Cohen Kadosh, Henik, & Rubinsten, 2008, Experiment 4). The earlier discussed discriminability mismatch in that study (as opposed to discriminability match in our study) cannot explain the different findings because in that study the size magnitude was easier to discriminate than that of the numerical magnitude which, according to Algom and Pansky (Algom et al., 1996; Pansky & Algom, 1999) predicts that numerical magnitude should not interfere with processing size magnitude. They observed the congruity effect despite the discriminability mismatch.

Interestingly, the fact that we do not observe a significant congruity effect in the physical size comparison task despite using exactly the same stimuli as in the numerical magnitude comparison task goes against the prediction of the decision stage conflict account (Santens & Verguts, 2011). Recall that according to that account, the congruity effect should be observed regardless of which exact dimension is task-relevant, as long as the decision alternatives in two tasks are exactly the same. It is also problematic for the shared magnitude code representations overlap because according to this account interference should arise whenever magnitudes in two dimensions are retrieved/computed regardless of which one is task-relevant and which one is task-irrelevant. Thus, both of these accounts have to somehow be modified in order to explain the lack of the congruity effect in the

physical size comparison task. We investigate the source of the congruity effect in the numerical comparison task before making conclusions.

In order to investigate whether the congruity effect that we observed in the numerical magnitude comparison task originates from the representational overlap at the level of magnitude codes or from a conflict at the decision stage, we conducted a follow-up experiment in which participants were asked to make a same/different judgment on the same stimuli.

4.3 Experiment 1c: same/different task with number words

We observed a significant difference between the congruent and the incongruent condition in the semantic comparison task with number words (Experiment 1a). Under the classical interpretation, this congruity effect arises from the overlapping magnitude code representations for the numerical magnitudes that are evoked by the number words and for the size magnitudes that are evoked by the font size difference. Therefore, this effect is seen as evidence in favor of number words evoking GMS representations. Under the alternative account, the congruity effect arises due to a conflict at the decision stage, simply because both the task-relevant and the task-irrelevant dimensions are processed in parallel. Since the response options are compatible for both of them, they subsequently compete for the response that should be given in case of the incongruent condition (e.g., the numerical magnitude dimension evokes a 'right larger' response, whereas the size magnitude dimension evokes a 'left larger' response.), but not in case of the congruent condition (e.g., both magnitude dimensions evoke a 'right larger' response). Under this interpretation, the congruity effect does not say anything about the interaction of GMS and numerical magnitudes conveyed by number words. It should be noted that while the congruity effect with number words has previously been interpreted as evidence in favor of the representational overlap account (Cohen Kadosh, Henik, & Rubinsten, 2008; Ito & Hatta, 2003), none of these previous studies have ruled out the decision stage conflict account. This is what we will look into now.

To tap into the origin of this congruity effect, we constructed a novel experiment where in part of the trials (the critical trials of the present experiment) the participants saw exactly the same stimuli as in the comparison task, but were asked to make a different decision: they had to decide whether the two presented words of a trial were repetitions of the same word or two different words.[14] In tri-

[14]Note that it is possible to construct two different versions of the same/different task parallel to the two versions of the comparison task that we had. In one version, the participants could be asked to decide whether two presented word are same while the match in size magnitude was manipulated. Alternatively, the participant could be asked to decide whether two presented font sizes were same while the numerical magnitude or scalar adjective meaning was manipulated.

als with the same stimuli as in the comparison task (Experiment 1a) two different number words were presented and thus, participants had to respond 'different' (and this held for both the congruent and the incongruent trials of Experiment 1a). We analyzed the reaction times for these trials. We added trials where participants saw the same number word on two sides of the screen (e.g., 'twee-twee' ['two-two'] or 'acht-acht' ['eight-eight']), and on these trials participants were supposed to respond 'same'. These trials were not analyzed.

The representational overlap account and the decision stage conflict account make different predictions for the 'different' trials of the same/different task. These predictions are illustrated in *Table 4.6* below. See the discussion of the same/different task under *Present study* for the reasoning behind these predictions.

Table 4.6: Predictions for differences between conditions under the representational overlap and decision stage conflict accounts of the size congruity effect for the comparison task with 'choose larger' decision polarity and for the same/different task. *'Left'* and *'right'* as well as *'same'* and *'different'* refer to response alternatives in the task.

	comparison task		**same/different task**	
	prediction under representational overlap	prediction under decision stage conflict	prediction under representational overlap	prediction under decision stage conflict
congruent: twee **acht**	magnitude code match	font: *right*, number: *right*	magnitude code match	font: *different*, number: *different*
incongruent: **twee** acht	magnitude code mismatch	font: *left*, number: *right*	magnitude code mismatch	font: *different*, number: *different*

A similar reasoning as the one we are using here in order to disentangle the two potential sources of this congruity effect has previously been applied in a study investigating subliminal priming of area size judgments by numerical magnitude, using Arabic digits (Lourenco et al., 2016, Experiment 2).

The setup of this same/different experiment is such that the participants can in principle give a fast response without ever processing the meaning of the number words that are presented, i.e., without accessing the numerical magnitude denoted by these number words (e.g., it should be straightforward to see that 'twee-acht' are two different words). Thus, in principle, if we do not observe a congruity effect in this task, it could be due to two reasons - either because there is no

We only administered the former version because we observed a congruity effect only in the semantic comparison tasks and were interested to investigate the source of specifically that observed congruity effect.

representational overlap between numerical magnitude and size magnitude or because the participants did not even activate the numerical magnitude of the number words (thus, not even allowing for any interaction between numerical and size magnitudes to take place). To our knowledge there is no published study that administered the same/different task in the specific stimulus set-up we are using (i.e., with number words that vary in their font size and congruity is manipulated). However, there has been a number of studies that administered same/different tasks with pairs of Arabic digits (Cohen et al., 2013; Ganor-Stern & Tzelgov, 2008; Wong & Szücs, 2013; Zhang, Xin, Feng, Chen, & Szűcs, 2017) or pairs of number words (Cohen et al., 2013; Dehaene & Akhavein, 1995; Wong, Bull, & Ansari, 2018) of the same font size with the goal to investigate specifically automaticity of activation of numerical magnitude information when it is not needed for the task. These studies used the so-called *numerical distance effect* as the signature of retrieval of the numerical magnitude. Specifically, we know that people can decide which of the numbers is larger/smaller when the two numbers are closer to each other (e.g., 4-5) than when they are further away from each other (e.g., 4-6 or 4-8, the larger the numerical difference the shorter observed reaction times; e.g., Buckley & Gillman, 1974; Hinrichs et al., 1981; Moyer & Landauer, 1967)[15]. To our knowledge, three studies so far investigated the numerical distance effect with two number words in the same/different task. The numerical distance effect has been observed in one of them (Dehaene & Akhavein, 1995) but not in the other two (Cohen et al., 2013; Wong et al., 2018). In one of the studies that did not observe a numerical distance effect with number words, instead an effect of physical similarity between number words has been observed (Cohen et al., 2013). In the present study, we look at the potential modulation of RTs by both the numerical distance and physical similarity.

If we do not observe a congruity effect and at the same time do not observe a numerical distance effect, then most likely participants have not even activated the numerical magnitude of the number words. On the other hand, if we do not observe a congruity effect, but do observe a numerical distance effect, then most likely participants *did* activate the numerical magnitude of the number words, but their representations did not overlap with those of size magnitude or size magnitude.

[15]However, the observed effect in the same/different task for Arabic digits has been recently shown to in fact originate from the physical similarity of digits themselves (similarity of digit symbols) rather than from processing their numerical magnitudes, so it has been suggested that participants in fact do not retrieve magnitudes of Arabic digits in the same/different task (e.g., Cohen, 2009; Wong & Szücs, 2013; Zhang et al., 2017).

4.3.1 Method

Participants

Because we had a restricted set of potential participants meeting the criteria in the pool of registered users of Prolific.ac, participation in this task was open to those who already completed the comparison task with scalar adjectives (Experiment 2a for which the data was collected at an earlier point in time). These participants have not seen number word stimuli before and have not completed a task requiring them to pay attention to the physical size of stimuli, so we did not expect them to be in any way different from completely naive participants.

Fifty-five participants completed the experiment. Three participants were excluded from the analysis because they read the first instructions in less than 10 s. Two further participants were excluded because they spent more than 30 minutes on the experiment. The mean age of the included participants was 25 years (SD 4.9; 33 male and 17 female). On average, they took approximately 14:43 minutes to complete the experiment (SD 02:37, min. 12, max. 25).

Stimuli

We used the same number word and font size combinations as for Experiments 1a and 1b to construct trials with an expected 'different' response. This means that we had five number word and font size combinations of interest as well as three filler combinations. Each number word in a pair appeared on both sides of the screen. In addition, we added trials with an expected 'same' response. Here, we presented the same number word on both sides of the screen albeit still in two different font sizes according to the font sizes that this number word was displayed in in the comparison tasks. This was done to keep these 'same' trials as similar as possible to the 'different' trials. For example, the pair 'twee-acht' was presented in font sizes 41 pt and 47 pt in 'different' trials (i.e., as in Experiments 1a and 1b). In addition, 'twee-twee' and 'acht-acht' were presented in font sizes 41 pt and 47 pt to create 'same' trials. Each font size appeared on both sides of the screen. The 'same' trials were not analyzed.

The proportion of 'different' and 'same' trials was 60:40 rather than balanced 50:50 similarly to the proportion that has been used in previous studies using the same-different paradigm (see Wong & Szücs, 2013, for reasoning for this choice).

Each participant saw 320 trials in total - 120 'different' trials of interest, 80 'same' trials with the same number word pairs as well as 120 filler trials. Each participant saw 60 trials of interest in the congruent condition in terms of magnitude (numerically larger/smaller number word presented in larger/smaller font size) and 60 trials of interest in the incongruent condition in terms of magnitude (numerically larger/smaller number word presented in smaller/larger font size). Because all of these were 'different' trials, comparison between these conditions could be made without a potential confound of the given response. The reaction

times of the 'same' trials were not compared to the 'different' trials since participants gave a different response here; we only provide the mean RT for this condition.

Procedure

The experimental procedure was identical to that of Experiments 1a and 1b except for instructions, response buttons, and the number of practice stimuli. Participants were instructed to indicate whether they saw the same word on both sides of the screen or different words. Half of the participants were told to press 'F' for the 'same' response and 'J' for the 'different' response in the first half of the experiment and vice versa for the second half of the experiment. The other half of participants received this response button mapping in the reversed order. Participants had a chance to practice both response mappings with feedback (at the beginning, when they read the first instructions and in the middle of the experiment, when the response button mapping was reversed). Because remembering the buttons for the same/different judgment might be more demanding than pressing a button on the side of the screen corresponding to a larger/smaller number, we included more practice items - ten items for each response button mapping.

4.3.2 Results

Participants included in the analysis made 3.42% errors in the whole experiment on average (min. 0%, max 9%). Data cleaning procedure resulted in exclusion of RTs of 4.63% (excluded incorrect responses are also counted here) of 'different' trials. The mean reaction time in the congruent condition was 714 ms (SD 175 ms, error rate: 2.7%), in the incongruent condition 719 ms (SD 182 ms, error rate: 2.2%) and in the 'same' trials it was 700 ms (SD 159 ms, error rate: 4.4%). Notice that the reaction times were overall somewhat faster for the 'same' decision than for the 'different' decision. These 'same' trials were not analyzed, so we now focus on the congruent and incongruent conditions within 'different' trials. Mean RTs and error rates per number word pair in each condition are given in *Table 4.7*.

The frequentist LME model with maximal random effect structure included a main effect of congruity and allowed for varying intercepts per-item and per-participant as well as varying slopes for the congruity effect in each case. This model did not converge. Exclusion of the varying slopes for the congruity effect per-participant resulted in a converging fit. The congruity effect was not significant ($\beta = 4, SE = 6.5, t = 0.62, p = 0.56$). The Bayesian LME model estimated for the congruity effect $\hat{\beta}$ = 3 ms, 95% CrI = [-17.04 25.56], BF_{01}=10.81; thus, there was strong evidence for the congruity effect being zero.

Table 4.7: Mean RT (SD), error rate per number word pair in 'different' trials in Experiment 1c, same/different task with number words.

number word pair	**congruent**	**incongruent**
'een-zes'	690 (159) ms, 2.34%	706 (169) ms, 1.33%
'twee-acht'	716 (174) ms, 2.50%	734 (185) ms, 2%
'twee-vijf'	694 (163) ms, 1.84%	703 (169) ms, 1.67%
'drie-acht'	754 (190) ms, 4.50%	736 (192) ms, 3.17%
'vier-acht'	720 (185) ms, 2.33%	720 (194) ms, 2.84%

To find out whether the participants actually processed numerical magnitude information in this set-up, we looked for the presence of the numerical distance effect in our data. The numerical distance was calculated as the difference between the larger and the smaller number within a pair. We also took into account the physical (i.e., visual) similarity of number words within each pair by calculating a score based on the confusion matrix of Geyer (1977) in parallel to how it was calculated by Cohen and colleagues in a previous same/different study with number words (Cohen et al., 2013; see supplemental online materials for the details about this calculation). The LME model with both factors included as predictors resulted in a significant effect of numerical distance ($\beta = 9, SE = 1.9, t = 4.95, p < 0.00001$) as well as significant effect of physical similarity ($\beta = -171, SE = 56.7, t = -3.02, p = 0.002$). Based on these results, we conclude that both physical similarity and numerical distance independently modulated the RTs. Thus, participants in this task did process numerical magnitude information.

To explore the data further, we looked at whether the participants perhaps learnt to ignore the numerical magnitude of the number words over the course of the experiment. To explore this possibility, we looked at the difference in the mean reaction times between the first half and the second half of the experiment.[16] Indeed, descriptively the difference in the mean RTs was somewhat larger in the first half of the experiment (congruent: 726 ms [SD 181], incongruent: 732 ms [SD 180]) than in the second half (congruent: 702 ms [SD 168], incogruent: 706 ms [SD 183]). Nonetheless, even in the first half of the experiment the congruity effect was not present. In the Bayesian model, there was moderate evidence for the interaction between the experiment half and congruity being zero ($\hat{\beta}$ = -6 ms, 95% CrI = [-28.3 14.27], BF_{01}=8.33). The frequentist LME models with a a reasonable random effect structure did not converge, so we do not report frequentist LME results here.

[16]Recall that response button mapping was counter-balanced across the participants in the both the first and second halves of the experiment, so response button mapping was not a confound here.

4.3.3 Interim discussion

We reasoned that if the congruity effect observed in the numerical comparison task (Experiment 1a) is explained purely by a conflict at the decision stage of processing, it should disappear when the response alternatives are such that they do not allow for such a conflict to arise. The data obtained with the same/different task indeed show that the congruity effect disappeared. We have in addition observed a significant modulation of the RTs by the numerical distance between the numerical magnitudes. We take this modulation as evidence that participants accessed the numerical magnitude information. We can thus rule out the possibility that the absence of a congruity effect is due to the fact that participants did some sort of superficial visual matching. Thus, overall we conclude that the congruity effect observed in the numerical comparison task with number words is likely to be driven by the conflict at the decision stage. Together with the lack of the congruity effect in the physical size comparison task with number words, this means that in the present study we do not observe any evidence for the recruitment of GMS during number word processing. Of course, it is possible that number words do recruit GMS, but that the tasks we used are not adequate for showing an involvement of GMS. We discuss these results in the wider context of research on number symbol processing in the *General Discussion* section.

Note that even if number word meaning does not interact with GMS representations, it is still possible that scalar adjectives' meaning does so given the differences between the properties of number symbols and scalar adjectives. Specifically, as we discussed in the *Introduction*, number symbols may not be compatible with GMS representations because they refer to exact, discrete quantities. In contrast, scalar adjectives do not refer to discrete magnitudes, so they are more compatible with GMS representations than number symbols are.

Even though the results of the present same/different task strongly suggest that the congruity effect in the size congruity paradigm originates at the decision stage, we still used this paradigm to look into processing of the scalar adjectives as well for several reasons. First, as discussed in the *Introduction*, there is evidence of this paradigm tapping into the interaction of magnitude codes of the task-relevant and task-irrelevant dimensions at least in case of Arabic digits and size magnitude (specifically, given the results of ERP studies showing both early and late interaction effects; Szucs & Soltesz, 2008). Second and more importantly, given that no previous study has looked at the possible interaction of scalar adjective meaning and GMS representations, comparing behavioral effects with scalar adjectives to those observed with number words would be a good starting point for this line of research. If we do observe a size congruity effect in case of scalar adjectives, we now know that (given the present results) it is likely to be at least partially driven by the presence of a conflict at the decision stage. It is, however, possible that a congruity effect for scalar adjectives is also partially driven by the representational overlap of scalar adjective meanings and size magnitude. Thus,

we will again need to investigate whether the congruity effect originates purely from a conflict at the decision stage in a same/different task.

4.4 Experiments 2a and 2b: comparison tasks with scalar adjectives

In the central experiments of this project, we use the reasoning and the experimental set-up of the size congruity paradigm to look at scalar adjectives and magnitude representations in GMS. In Experiments 2a and 2b, participants saw pairs of scalar adjectives on the screen and were asked to make a decision about their meaning (Experiment 2a) or about the font size in which they were presented (Experiment 2b). The same stimuli were presented in both experiments. The experimental design, procedures and number of trials for these experiments were identical to Experiments 1a and 1b with number words.

In the semantic comparison task with scalar adjectives (Experiment 2a), we employed a novel task. The participants saw pairs of antonymous scalar adjectives referring to continuous property dimensions (e.g.,'kort-lang' ['short-long'], 'laag-hoog' ['low-high'], 'licht-zwaar' ['light-heavy']) in different font size combinations. Note that we did not use the adjective pairs that would be used to describe the physical size contrast itself, i.e. 'large-small'. Participants were asked to indicate the adjective in the pair that 'means more/less of something', and they were given several examples such as 'young-old' where 'old' refers to *more* in terms of age. Thus, for this task participants had to understand the dimension which the adjectives describe, and they had to decide which of the adjectives refers to more/less on this particular dimension. The exact instructions and examples given to the participants are provided in the *Procedure* section below. The low error rate that we observed (between 0.83-7%) demonstrates that the participants did not have any difficulty with this task. The physical size comparison task with scalar adjectives (Experiment 2b) was the same as for number words (Experiment 1b) - participants were asked to choose the word that was printed in the larger/smaller font size. They received same instructions as participants in physical size comparison with number words (Experiment 1b).

As for the experiments with number words, we attempted to match the two dimensions (meaning and physical size magnitude) in both variability and discriminability as much as possible. However, despite these attempts, it was not possible to match stimuli in terms of discriminability - the general speed for processing adjective meaning was slower than that for processing font size. This was the case for the following reason. Discriminability of stimuli in the task-relevant and task-irrelevant dimensions is matched by matching the reaction times of the judgments in each of the dimensions separately. The reaction times depend on the difficulty of the comparison in the respective task-relevant dimension. There is no way to manipulate the speed of the symbolic meaning dimension, so this

speed is a given fact. By contrast, we can manipulate the physical size judgment - for example, we can make it slower by making the font size difference smaller. However, making the font size differences increasingly smaller also leads to a larger number of errors which is undesirable because we would like to have a roughly comparable error rate in the two dimensions as well. As a result of this, it was not possible to match discriminability fully. As an alternative to full matching, we opted for choosing font size combinations that were closest to the symbolic judgment RTs for the individual specific adjectives pairs.

Let us consider how the difference discriminability in the two dimensions can affect our results. The reaction times for the physical size comparison of font sizes used in the present study was on average 87 ms shorter than the reaction time for the semantic comparison of the adjective pairs (see *Table 4.8*). According to the findings of Algom and Pansky (Algom et al., 1996; Pansky & Algom, 1999), the fact that the discriminability of physical size is easier than that of scalar adjective pairs means that physical size will attract more attentional resources than the numerical magnitude. For semantic comparison task this means that the congruity effect may arise simply because physical size is more salient and attracts the attentional resources. If this is the case, the congruity effect should be different for each of the stimulus pairs depending on how large the difference in discriminability between the task-relevant and task-irrelevant dimensions is. We consider this possibility in the *Interim discussion* of these experiments below. For the physical size comparison task the discriminability difference means that the adjective meaning may not be able to interfere with the physical size comparison simply because it will not be able to attract enough attentional resources (i.e., it may not be processed fast enough). However, recall that Cohen Kadosh and colleagues (2008) have observed a significant congruity effect in a physical size comparison task with number words despite a larger mismatch in discriminability. Thus, possibly in case of our physical size comparison task we will still be able to observe a congruity effect (especially given that the mismatch in discriminability is in fact smaller than in the study of Cohen Kadosh and colleagues where it was 100-300 ms). Nonetheless, if we do not observe a congruity effect, we will not be able to completely rule out the possibility that discriminability difference did not allow for this effect to emerge.

4.4.1 Method

Participants

Fifty-nine participants completed Experiment 2a, i.e., the semantic comparison task with scalar adjectives. Six participants were excluded from the analysis because they gave incorrect responses in more than 15% of trials. Two participants were excluded because they read the first instructions in less than 10 s. One further participant was excluded because they spent more than 30 minutes on

the experiment. The mean age of the included participants was 25 years (SD 4.9; 33 male and 17 female). On average, they took approximately 15:01 minutes to complete the experiment (SD 01:55, min. 12, max. 20).

Sixty-three participants completed Experiment 2b, i.e., the physical size comparison task with scalar adjectives. Eleven participants were excluded from the analysis because they gave incorrect responses in more than 15% of trials. One participant was excluded because the data for half of the trials was lost due to technical reasons. Finally, one participant was excluded because they spent more than 30 minutes on the experiment. The mean age of the included participants was 23 years (SD 4.2; 30 male and 20 female). On average, they took approximately 15:24 minutes to complete the experiment (SD 02:24, min. 12, max. 24).

Stimuli

The configuration of stimuli, number of trials of interest and filler trials was parallel to the ones described for Experiments 1a and 1b with number words. Here and in the rest of the methods section, we only mention the differences from the methods described for number words.

We used five pairs of scalar adjective pairs: 'kort-lang' ['short-long'], 'laag-hoog' ['low-high'], 'licht-zwaar' ['light-heavy'], 'dun-dik' ['thin-thick'], and 'stil-luid' ['quiet-loud'], presented in five combinations of font sizes, respectively: 43-48 pt, 41-47 pt, 37-42 pt, 38-42 pt, 41-46 pt. As for the number words, we were again restricted in the number of suitable scalar adjective pairs because we matched the adjectives within a pair in terms of number of letters in order to avoid any length confounds. In order to match the task dimensions on discriminability, we collected data in a norming study prior to the experiments from the same 30 participants that completed the norming for number words (none of these participants took part in the actual experiments; see supplemental online materials for details on this norming study). Because the participants were in general considerably slower on judgments of adjective meanings than on font size judgments, it was not possible to fully match the scalar adjective pairs with font size combinations in terms of reaction times. Instead, we chose font size combinations that were closest to the adjectives pairs in terms of RTs. The mean RTs and error rates observed for the selected scalar adjective and font size pairs are provided in *Table 4.8*.

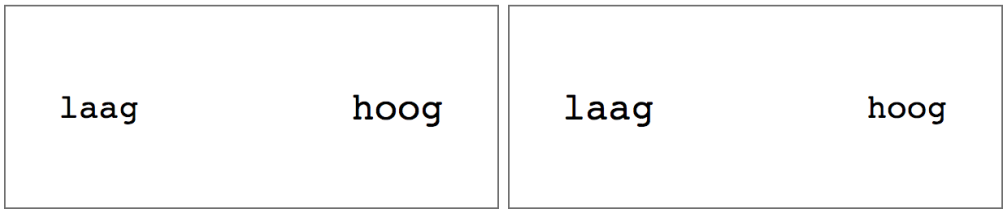
The five scalar adjective pairs of interest were intermixed with three further filler pairs: 'weinig-veel' ['few/little-many/much'], 'smal-breed' ['narrow-broad'], and 'langzaam-snel' ['slow-fast'] presented in font sizes 42-46 pt, 38-43 pt, 38-44 pt respectively. In case of filler trials, the discriminability was not matched.

Examples of displays in congruent and incongruent conditions are shown in *Figure 4.3*

Table 4.8: Mean RT (SD) and error rate observed in stimuli pre-test for the selected pairs of scalar adjective and font size combinations.

adjective pair	**RT (SD)**	**error rate**	**font size pair**	**RT (SD)**	**error rate**
'kort-lang'	887 (208) ms	7%	43-48 pt	787 (235) ms	6.97%
'laag-hoog'	793 (206) ms	3.83%	41-47 pt	720 (222) ms	2.43%
'licht-zwaar'	834 (229) ms	0.83%	37-42 pt	747 (229) ms	1.74%
'dun-dik'	829 (217) ms	5.33%	38-42 pt	761 (254) ms	5.92%
'stil-luid'	884 (246) ms	4.83%	41-46 pt	774 (222) ms	5.56%
across all pairs	844 (224) ms	4.36%	across all pairs	757 (233) ms	4.52%

Figure 4.3: Examples of displays in congruent and incongruent conditions in Experiments 2a and 2b.



(a) congruent condition (b) incongruent condition

Procedure

The experimental procedure was identical to that of the Experiments 1a and 1b except for instructions. The following instructions were given to the participants who completed semantic comparison task: "In this experiment, you will see two words at the center of the screen. Your task is to indicate which of the two words means *more of something* by pressing the corresponding key on your keyboard. This means more in terms of what the meaning of the word refers to. For example, in the pair *'old'* and *'young'*, 'old' is more in terms of age. In the pair *'expensive'* and *'cheap'*, 'expensive' is more in terms of price. In the pair *'a lot'* and *'one'*, 'a lot' is more in terms of quantity. If the word that means more is on the right side, press 'P' and if it is on the left side, press 'Q'. For example, you may see 'old young'. In this case, the word which means more is 'old', on the left side, and you should press 'Q'. If it was 'young old', then you would need to press 'P'. You should only compare the meaning of the two words and ignore the other properties. [..]". For the 'choose less' decision polarity, the instructions were identical except 'more' was substituted by 'less' and examples were adjusted. In the practice trials, participants saw the adjective pairs given as examples in the instructions intermixed with other pairs (e.g., 'full-empty', 'fat-slim', etc.).[17] None of the adjectives that appeared as an example or in practice trials appeared

[17]See supplemental online materials for full instructions in Dutch as well as all practice trials.

in the experimental trials.

The texts of the instructions given to the participants who completed physical size comparison with scalar adjectives were identical to those given in the physical size comparison with number words (Experiment 1b). Participants were instructed to indicate the side of the screen with a word in larger or smaller font size. Examples and practice trials given to participants in Experiment 2b were different, however. Here, they saw some scalar adjective pairs (e.g., 'old-young', 'full-empty') and some color adjective pairs (e.g., 'red-blue') as an example and in the practice trials. None of the adjectives that appeared as an example or in a practice trial appeared in the experimental trials.

4.4.2 Results

In Experiment 2a, i.e., the semantic comparison task with scalar adjectives, participants included in the analysis made 5.16% errors in the whole experiment on average (min. 0%, max 13%). Data cleaning procedure resulted in exclusion of RTs of 6.6% of trials of interest (excluded incorrect responses are also counted here). The resulting mean RTs and error rates per congruity overall and in each decision polarity are given in *Table 4.9* and visually depicted in *Figure 4.4a*. Mean RTs and error rates per adjective pair across decision polarities are given in *Table 4.10*.

The model with maximal random effect structure for Experiment 2a did not converge. The random effect structure was gradually simplified to achieve a converging non-singular model fit. The final model included a varying intercept per-item as well as varying intercept per-participant allowing for varying slopes for the effect of decision polarity. In this model, the main effect of congruity was significant ($\beta = 16, SE = 5.7, t = 2.82, p = 0.005$) along with the main effect of decision polarity ($\beta = 28, SE = 11.9, t = 2.35, p = 0.022$). The interaction of congruity and decision polarity was not significant ($\beta = -8, SE = 8.1, t = -1.05, p = 0.28$).

The Bayesian LME model estimated for the main effect of congruity $\hat{\beta} = 19$ ms, 95% CrI = [7.71 31.86], BF_{10}=9.09; for the main effect of decision polarity $\hat{\beta} =$ 24 ms, 95% CrI = [-14.1 59.95], BF_{01}=1.86, for the interaction between congruity and decision polarity $\hat{\beta}$ = -6 ms, 95% CrI = [-25.85 13.1], BF_{01}=8.94. Thus, there is moderate evidence that the congruity effect is not zero, no clear evidence for or against the decision polarity effect being zero and moderate evidence that the interaction between congruity and decision polarity is zero.

In Experiment 2b, i.e., the physical size comparison task with scalar adjectives, participants included in the analysis made 7.08% errors in the whole experiment on average (min. 2%, max 14%). Data cleaning procedure resulted in exclusion of RTs of 6.59% of trials of interest (excluded incorrect responses are also counted

Table 4.9: Mean RT (SD), error rate overall and for each decision polarity in Experiment 2a, semantic comparison with scalar adjectives.

decision polarity	**congruent**	**incongruent**
overall	846 (228) ms, 5.04%	858 (219) ms, 5.48%
'choose more'	833 (220) ms, 5.12%	848 (212) ms, 5.24%
'choose less'	861 (236) ms, 4.96%	869 (226) ms, 5.72%

Table 4.10: Mean RT (SD), error rate per adjective pair (both decision polarities) in Experiment 2a, semantic comparison with scalar adjectives.

adjective pair	**congruent**	**incongruent**
'kort-lang'	889 (239) ms, 7.8%	903 (231) ms, 9.7%
'laag-hoog'	806 (210) ms, 3.8%	825 (202) ms, 4.3%
'licht-zwaar'	805 (212) ms, 2.0%	815 (201) ms, 2.0%
'dun-dik'	853 (228) ms, 6.5%	864 (231) ms, 6.5%
'stil-luid'	887 (238) ms, 5.11%	891 (217) ms, 4.9%

here). The resulting mean RTs and error rates per congruity overall and in each decision polarity are given in *Table 4.11* and visually depicted in *Figure 4.4b*. Mean RTs and error rates per adjective pair across decision polarities are given in *Table 4.12*.

The model with maximal random effect structure for Experiment 2b resulted in a singular fit. It was not possible to achieve a non-singular converging fit without drastically simplifying the random effect structure (which we believe would not be justified in our case since we know there must be some variability by-participant and by-item). For this reason, we examined the fit of the model with maximal random structure even though it resulted in a singular fit.[18] None of the effects were significant (main effect of congruity - $\beta = 30, SE = 50.7, t = 0.60, p = 0.58$; main effect of decision polarity - $\beta = 36, SE = 18.0, t = 2.01, p = 0.061$; interaction of congruity and decision polarity - $\beta = 0.8, SE = 10.3, t = 0.07, p = 93$). The pattern that we observed here is parallel to the one observed in Experiment 1b. Whereas the mean reaction times in the congruent and congruent conditions differ in the expected direction, this difference is not consistently present for each of the adjective pairs (as can be seen in *Table 4.12*). This is reflected in a non-significant effect in the LME model.

The Bayesian LME model estimated for the main effect of congruity $\hat{\beta} = 18$

[18]The singular fit means that variances of one or more linear combinations of effects were estimated to be (close to) zero. However, we have good reasons for a maximal random effect structure - we know that there must be at least some variation in the size of all effects per participant and per item. In short, we believe that estimates from the maximal random effect structure models can still be informative in case of our experiments even if they result in a singular fit.

ms, 95% CrI = [-44.63 79.09], BF_{01}=2.8; for the main effect of decision polarity $\hat{\beta}$ = 20 ms, 95% CrI = [-14.66 54.53], BF_{01}=2.49, for the interaction between congruity and decision polarity $\hat{\beta}$ = 1 ms, 95% CrI = [-17.83 20.39], BF_{01}=13.34. Thus, there is no clear evidence for or against the congruity effect being zero, no clear evidence for or against the decision polarity effect being zero and strong evidence that the interaction between congruity and decision polarity is zero.

Table 4.11: Mean RT (SD), error rate overall and for each decision polarity in Experiment 2b, physical size comparison with scalar adjectives.

decision polarity	**congruent**	**incongruent**
overall	761 (238) ms, 3.39%	787 (248) ms, 7.56%
'choose larger'	744 (232) ms, 2.78%	769 (246) ms, 7.21%
'choose smaller'	779 (244) ms, 4.00%	806 (250) ms, 7.92%

Table 4.12: Mean RT (SD), error rate per adjective pair (both decision polarities) in Experiment 2b, physical size comparison with scalar adjectives.

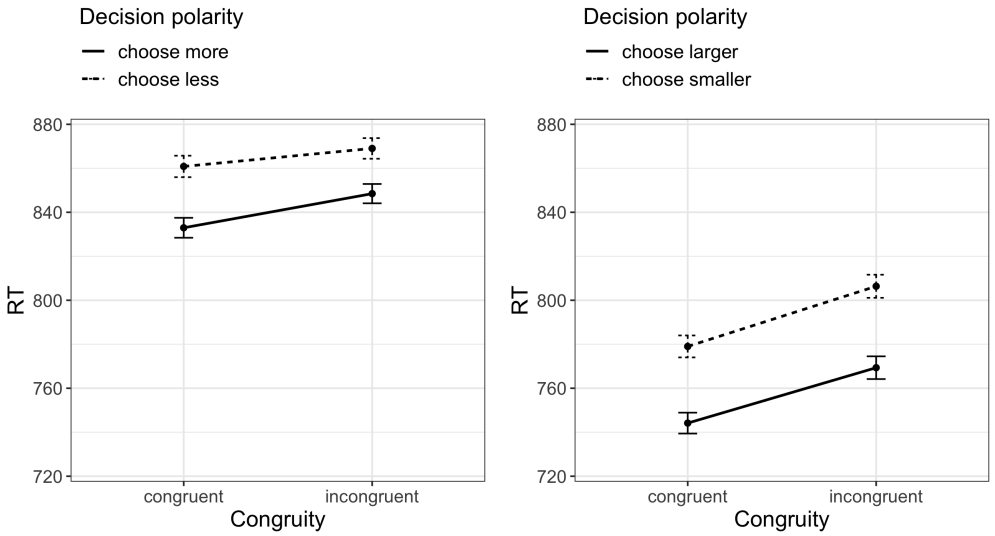
adjective pair	**congruent**	**incongruent**
'kort-lang'	788 (232) ms, 3.71%	789 (256) ms, 5.82%
'laag-hoog'	730 (229) ms, 1.31%	768 (219) ms, 3.01%
'licht-zwaar'	809 (275) ms, 7.72%	700 (189) ms, 1.91%
'dun-dik'	714 (207) ms, 1.41%	896 (299) ms, 21.29 %
'stil-luid'	771 (235) ms, 2.81%	812 (244) ms, 5.82%

4.4.3 Interim discussion

The pattern of effects that we observe in Experiments 2a and 2b is parallel to what we observed for number words (Experiment 1a and 1b). In the semantic comparison task, we observed a congruity effect. This congruity effect was present for both decision polarities (there was moderate evidence that the effect of interaction of decision polarity and congruity is zero).

Before we investigate its source in a same/different task, we need to consider the alternative explanation in terms of discriminability differences. Recall that in Experiments 2a and 2b we were able to match the stimuli in terms of variability, but not in terms of discriminability. It is possible that the size magnitude interfered with numerical magnitude processing simply because it was more salient. If this was the case, the adjective pairs with a clearer discriminability should have resulted in a larger congruity effect. However, this does not seem to be the case

Figure 4.4: Mean RTs per congruity and decision polarity in comparison tasks with scalar adjectives. The error bars depict the standard error value.



(a) Experiment 2a: semantic comparison task with scalar adjectives.

(b) Experiment 2b: physical size comparison task with scalar adjectives.

when inspecting the means informally: the adjective pair with the largest difference in discriminability ('stil-luid', size comparison 110 ms faster than adjective comparison) resulted in the smallest congruity effect (the difference in the means just 4 ms). To explore this possibility formally, we re-ran the converging non-singular frequentist LME model described above additionally including the main effect of discriminability difference as well as interaction between discriminability difference and congruity as fixed factors. This model did not result in a significant main effect of discriminability or in an interaction between discriminability and congruity (though the congruity effect was also non-significant in this model so discriminability did explain some variance that was previously attributed to congruity). In addition, this model did not result in a better fit to the data than the original one ($\chi^2(2) = 2.94, p = 0.22$). We interpret these results as showing that discriminability difference does not clearly modulate the congruity effect. We, therefore, conclude that while discriminability difference remains a plausible reason behind part of the the observed congruity effect, it is unlikely to completely account for it.

For the physical size comparison task with scalar adjectives we did not observe a congruity effect. In fact, the evidence is inconclusive (BF_{01}=2.8) though the null hypothesis is supported by the data slightly more than the alternative hypothesis.

4.5 Experiments 2c: same/different task with scalar adjectives

The same/different task with scalar adjectives was parallel to the same/different task with number words (Experiment 1c, above).

4.5.1 Method

Participants

Because we had a restricted set of available participants meeting the criteria in the pool of registered users of Prolific.ac, participation in this task was open to those who had participated in the comparison task with number words (Experiment 1a for which the data was collected at an earlier point in time). These participants have not seen scalar adjective stimuli before and have not completed a task requiring them to pay attention to the physical size of stimuli, so we did not expect them to be in any way different from completely naive participants (as already discussed for participants of Experiment 1c above).

Fifty-five participants completed the experiment. Three participants were excluded from the analysis because they read the first instructions in less than 10 s. Two participants were excluded because they gave incorrect responses in more than 15% of trials. One further participant was excluded because they spent more than 30 minutes on the experiment. Due to a miscalculation at the data collection stage, we ended up with only 49 participants with valid datasets in this experiment instead of the planned 50. The mean age of the included participants was 25 years (SD 4.7; 32 male and 17 female). On average, they took approximately 15:05 minutes to complete the experiment (SD 02:57, min. 11, max. 29).

Stimuli

We constructed stimuli in a parallel way to how it is described for Experiment 1c, same/different task with number words, but used the adjective stimuli of Experiments 2a and 2b.

Procedure

The experimental procedure was identical to the one for Experiment 1c, same/different task with number words. The only difference was that the example items in the instructions and the practice items consisted of adjectives instead of number words.

4.5.2 Results

Participants included in the analysis made 3.81% errors in the whole experiment on average (min. 1%, max 10%). Data cleaning procedure resulted in exclusion of RTs of 5.24% (excluded incorrect responses are also counted here) of congruent and incongruent trials according to the representational overlap account, i.e., in 'different' trials. The mean reaction time in the congruent condition was 698 ms (SD 163 ms, error rate: 2.8%), in the incongruent condition 710 ms (SD 174 ms, error rate: 3.7%) and in the 'same' trials it was 693 ms (SD 158 ms, error rate: 6.1%). The 'same' trials were not analyzed. Mean RTs and error rates per number word pair in the congruent and incongruent 'different' trials are given in *Table 4.13*.

Table 4.13: Mean RT (SD), error rate per scalar adjective pair in 'different' trials in Experiment 2c, same/different task with scalar adjectives.

adjective pair	**congruent**	**incongruent**
'kort-lang'	691 (158) ms, 1.54%	683 (165) ms, 3.24%
'laag-hoog'	702 (158) ms, 3.26%	718 (175) ms, 3.40%
'licht-zwaar'	676 (174) ms, 1.54%	682 (173) ms, 0.86%
'dun-dik'	714 (159) ms, 5.64%	726 (175) ms, 4.79%
'stil-luid'	709 (166) ms, 2.40%	747 (174) ms, 6.51%

The frequentist LME model with maximal random effect structure included a main effect of congruity and allowed for varying intercepts per-item and per-participant as well as varying slopes for the congruity effect in each case. This model did not converge. The model excluding varying slopes for the congruity effect per participant converged. The congruity effect was not significant ($\beta = 12, SE = 7.3, t = 1.76, p = 0.15$). The Bayesian LME model estimated for the congruity effect $\hat{\beta} = 5$ ms, 95% CrI = [-14.95 24.52], BF_{01}=9.5; thus, there was moderate evidence for the congruity effect being zero.

In parallel to the exploratory analysis for the same/different data with number words (Experiment 1c), here we again explored whether there was learning effect over the course of the experiment by comparing mean reaction times in trials shown in the first as opposed to the second half of the experiment. Again, descriptively the RTs did get shorter over the course of the experiment, and the difference in the mean RTs between congruent and incongruent conditions was somewhat larger in the first half of the experiment (congruent: 706 ms [SD 163], incongruent: 726 ms [SD 182]) than in the second half of the experiment (congruent: 690 ms [SD 163], incongruent: 695 ms [SD 164]). The frequentist LME models with a a reasonable random effect structure did not converge, so we do not report frequentist LME results here. The Bayesian model showed moderate evidence that the interaction between congruity and experiment half was zero ($\hat{\beta}$

= -4 ms, 95% CrI = [-29.04 20.18], BF_{01}=9.38). Thus, even in the first half of the experiment there was no congruity effect.

4.5.3 Interim discussion

Whereas we observed a congruity effect in the size congruity paradigm with scalar adjectives (Experiment 2a), the data from the present same/different task did not show a significant congruity effect. We found moderate evidence for the congruity effect being absent in the same/different task. Thus, we conclude that the congruity effect in the semantic comparison task with scalar adjectives was likely due to the conflict at the decision stage of processing. Combined with the results of the physical size comparison task with scalar adjectives (Experiment 2b), the present series of experiments does not show evidence for recruitment of GMS representations during the processing of scalar adjectives.

4.6 General discussion

In the present project, we put forward the hypothesis that scalar adjectives such as 'tall', 'short', 'long', 'big', 'loud', etc. are symbolic references to generalized magnitude system representations, and that our language comprehension system recruits GMS representations when processing these adjectives. Consistent with the observed properties of the representation format of GMS, scalar adjectives refer to only approximate values and their applicability as descriptions of magnitude depends on relative rather than absolute values. While it has been suggested in the past that processing numerals (e.g., Arabic digits or number words) recruit GMS representations, as far as we know, no research has previously looked at the potential connection of scalar adjectives and GMS representations. We compared processing of scalar adjectives to processing of number words because number words are similar to scalar adjectives in their reference to magnitude information, and the relationship of number words to GMS has previously been investigated.

In Experiments 1a and 1b, we used the size congruity paradigm with number words. The size congruity paradigm has been used in the past to look at the interaction of magnitudes evoked by number symbols and by physical size magnitude. We observed a clear congruity effect in a semantic comparison task (i.e., where numerical magnitude was the task-relevant and physical size magnitude was the task-irrelevant dimension, Experiment 1a). There was no significant congruity effect when in the task-relevant and task-irrelevant dimensions were reversed (Experiment 1b). In Experiment 1c, we used the same/different paradigm with the same stimuli as in Experiments 1a and 1b. This paradigm allows to eliminate the potential conflict at the decision stage of processing as the origin of the congruity effect that we observed in the semantic comparison task. With the same/different task, we no longer observed a significant congruity effect; in

fact, we had strong statistical evidence that there was no hints of a congruity effect. We thus conclude that the congruity effect that we observed with number words in the numerical comparison task was presumably primarily driven by the response conflict at the decision stage. In summary, the experiments on number words do not provide evidence that the comparison of numerical magnitudes carried by number words recruits GMS representations. The implications of this result for research on number symbol processing will be discussed below.

The reasoning behind and the design of Experiments 2a-c on scalar adjectives were parallel to the experiments 1a to 1c on the number words. Here, we again observed a congruity effect only in the semantic comparison task (participants compared the meaning of antonymous pairs of scalar adjectives and physical size magnitude was the task-irrelevant dimension), and again a congruity effect was no longer present in the same/different task with the same stimuli. Thus, as for number words, the congruity effect was primarily driven by a response conflict in the decision stage. The results of the present series of experiments thus do not provide support for the hypothesis that GMS is recruited in the processing of scalar adjective. This in turn either means that GMS is not involved in processing of scalar adjectives at all, or it implies that the size congruity paradigm is not suited to demonstrate the involvement of GMS in the processing of scalar adjectives.

4.6.1 Implications of the present results for number symbol processing

As discussed in the *Introduction*, to our knowledge, the size congruity paradigm has previously been used to look at number word processing (i.e. not digits) in only three studies (Cohen Kadosh, Henik, & Rubinsten, 2008; Foltz et al., 1984; Ito & Hatta, 2003). A significant congruity effect in the numerical comparison task has been observed in all three studies and, in line with these studies, also in Experiment 1a of the present study. Our results thus add to this evidence in a new language - Dutch. In addition, the stimuli in our study were matched in terms of discriminability in the task-relevant and task-irrelevant dimensions. Finally, our study was the first one to administer the 'choose smaller' decision polarity with number words, and we show that the congruity effect is identical for this decision polarity. Recall that the exact locus of the congruity effect in the numerical comparison task with number words has not been addressed in the past. The results of the same/different task (Experiment 2c) strongly suggest that the congruity effect is primarily driven by a response conflict in a decision stage. This implies that the the decision stage could also have been the primary source for the congruity effects in these past studies as well.[19]

[19]Note, however, that the account of the conflict at the decision stage by Santens and colleagues (Santens & Verguts, 2011) that we have discussed in the *Introduction* cannot fully

Two previous studies that administered a physical size comparison task with number words did not observe a significant congruity effect (in line with our Experiment 2b) while observing a congruity effect in the same task with Arabic digits or Kanji numerals (Kanji is an ideographic script; (Ito & Hatta, 2003; Cohen Kadosh, Henik, & Rubinsten, 2008, Experiment 1) or observed what appeared to be a qualitatively different congruity effect (Cohen Kadosh, Henik, & Rubinsten, 2008, Experiment 4). Ito and Hatta interpret their results as suggesting that number words do not have a strong automatic connection to the numerical magnitude representations in general or at least in case of a 'less attentive processing condition' (such as when the numerical magnitude is the task-irrelevant dimension). Cohen Kadosh and colleagues go even further and argue that our cognitive system has separate comparison mechanisms for number words and Arabic numbers, and that the numerical magnitude representations of the two notations are potentially distinct, though highly interconnected. These previous studies, however, did not fully match the discriminability and variability of the stimuli (Algom et al., 1996; Pansky & Algom, 1999) along the numerical and physical size dimensions. In our physical size comparison task, the discriminability and variability *were* matched, but we still failed to observe a congruity effect. The evidence for or against the congruity effect in this task was inconclusive, but still the data was more consistent with the possibility that the congruity effect was zero (BF_{01}=2.2). While we did not investigate Arabic digits in our own study, the results of the present experiment are consistent with the interpretations of Ito and Hatta and Cohen Kadosh and colleagues.

In summary, the present study shows that the size congruity paradigm does not provide support for the hypothesis that number words recruit GMS representations.

Given this conclusion, let us now consider how informative the results from the size congruity paradigm can be regarding recruitment of GMS representations for Arabic digits (rather than number words). First, in fact the pattern of congruity effects with Arabic digits has been observed to be different from that for number words. As discussed earlier, for Arabic digits, congruity is observed for both numerical comparison and physical size comparison tasks and is differently modulated by the numerical distance (Cohen Kadosh, Henik, & Rubinsten, 2008; Ito & Hatta, 2003). Furthermore, the congruity effect observed with Arabic digits has in fact been observed to be modulated by the decision polarity - the congruity effect is larger for the 'choose larger' instruction than for the 'choose

account for the results we observed in the present study either. That is because it suggests that a congruity effect should be observed regardless of which exact dimension is task-relevant and which is task-irrelevant, as long as they have two compatible response options that can compete with each other. If that were the case, we should have observed a congruity effect in the physical size comparison task as well, but we in fact did not observe a congruity effect there. Thus, additional assumptions are needed to account for the full range of the observations (we will not provide a solution here).

smaller' instruction (Arend & Henik, 2015; Tzelgov et al., 1992). The presence of a modulation of the congruity effect by decision polarity is usually used as an argument against an explanation of the congruity effect originating in a conflict at the decision stage (Arend & Henik, 2015). Finally, EEG studies on the size congruity effect with Arabic digits detected ERP signatures of the congruity effect at an early point after stimulus presentation (150-250 ms after onset) which can be seen as the stage when the magnitude representations are retrieved as well as a later stage which can be seen a as decision stage conflict (Szucs & Soltesz, 2008). Taken together, there is thus still good evidence that in case of Arabic digits the size congruity effect suggests some recruitment of GMS representations, i.e., that the congruity effect results from the overlap in the magnitude representations for the numerical magnitude and size magnitude.

However, suppose that it turns out that the size congruity effect for Arabic digits is not driven by the involvement of GMS, do we then have to conclude that GMS is not involved in the processing of Arabic digits either? Evidence for interaction of number symbols and GMS representations has recently been observed in a different experimental paradigm. Lourenco and colleagues (2016) observed subliminal priming effects from Arabic digits to judgments of area. In the most interesting experiment of this study (Experiment 2), two digits were presented as primes. In the critical condition, one of the prime digits was presented in white color and the other one in black color. The digits were followed by the presentation of an array of white and black rectangles. Participants were asked to indicate whether the addition of the areas (i.e., the cumulative area) of the white rectangles and that of the black rectangles was 'same' or 'different'. In 'different' trials, the color of the numerically larger prime digit could correspond to or be different from the color of the rectangles with larger area. They observed shorter RTs when the color of the numerically larger digit matched the color of the rectangles with the larger cumulative area in 'different' trials. Because response alternatives were not aligned with the magnitude match or mismatch between primes and cumulative areas, this design excluded the possibility that decision stage conflict is responsible for this congruity effect. This study thus provides evidence for the involvement of GMS in the processing of number symbols and GMS for which the criticism directed at the size congruity paradigm does not apply.

4.6.2 Implications of the present results for scalar adjective processing

The central goal of the present project was to present and evaluate the hypothesis that the human language processing system makes use of the generalized magnitude system representations during the retrieval of the meaning of scalar adjectives and the construction of a mental model of the communicated informa-

tion. Like number symbols, scalar adjectives could also be symbolic references to magnitude information and share a number of features with the GMS representations. We tested this hypothesis by looking at a potential interference between retrieval of scalar adjective meaning and computation of physical size magnitude carried out simultaneously. The data collected in the present project do not support this hypothesis.

Although this series of experiments did not provide support for our hypothesis, the hypothesis and its variations remain interesting and can be investigated in a number of other ways in future. These are discussed in the *Summary and avenues for future research* chapter of the present thesis.

4.7 Conclusion

The present series of experiments investigated processing of number words and of scalar adjectives both of which can be seen as symbolic references to magnitude information. We investigated whether processing of these lexical items recruits generalized magnitude system representations. The data collected for number words add to the existing literature by showing that the congruity effect with number words is most likely driven by a response conflict at the decision stage of processing. A similar conclusion was reached for the processing of scalar adjectives in the size congruity paradigm. The results of the present study thus do not support the hypothesis that the processing of scalar adjectives involves GMS representations. Furthermore, the present study reveals some serious limits of the size congruity paradigm for studying the representations involved in magnitude processing as the results strongly suggest that response conflicts in a late decision stage have a strong influence on the results.

4.8 Data Accessibility

Stimuli, raw data, data collection and analysis scripts are available on

https://osf.io/kh6eb/.

Chapter 5

Semantic and syntactic composition of minimal adjective-noun phrases in Dutch: an MEG study

Abstract

[1]The possibility to combine smaller units of meaning (e.g., words) to create new and more complex meanings (e.g., phrases and sentences) is a fundamental feature of human language. In the present project, we investigated how the brain supports the semantic and syntactic composition of two-word adjective-noun phrases in Dutch, using magnetoencephalography (MEG). The present investigation followed up on previous studies reporting a composition effect in the left anterior temporal lobe (LATL) when comparing neural activity at nouns combined with adjectives, as opposed to nouns in a non-compositional context. The first aim of the present study was to investigate whether this effect, as well as its modulation by noun specificity and adjective class, can also be observed in Dutch. A second aim was to investigate to what extent these effects may be driven by syntactic composition rather than primarily by semantic composition as was previously proposed. To this end, a novel condition was administered in which participants saw nouns combined with pseudowords lacking meaning but agreeing with the nouns in terms of grammatical gender, as real adjectives would. We failed to observe a composition effect or its modulation in both a confirmatory analysis (focused on the cortical region and time-window where it has previously been reported) and in exploratory analyses (where we tested multiple regions and an extended potential time-window of the effect). A syntactically driven

[1]This chapter is based on: Kochari, A., Lewis, A., Schoffelen, J. M., & Schriefers, H. Semantic and syntactic composition of minimal adjective-noun phrases in Dutch: an MEG study. *Manuscript.*

composition effect was also not observed in our data. We do, however, successfully observe an independent, previously reported effect on single word processing in our data, confirming that our MEG data processing pipeline does meaningfully capture language processing activity by the brain. The failure to observe the composition effect in LATL is surprising given that it has been previously reported in multiple studies. Reviewing all previous studies investigating this effect, we propose that materials and a task involving imagery might be necessary for this effect to be observed. In addition, we identified substantial variability in the regions of interest analysed in previous studies, which warrants additional checks of robustness of the effect. Further research should identify limits and conditions under which this effect can be observed. The failure to observe specifically a syntactic composition effect in such minimal phrases is less surprising given that it has not been previously reported in MEG data.

5.1 Introduction

One of the fundamental properties of human language is compositionality - the possibility to combine smaller units of meaning into larger ones creating new, integrated meanings. This is possible at multiple levels; for example, morphemes are combined into words, words are combined into phrases, phrases are combined into sentences. Here, we focus on how our brains combine individual words into phrases - for example, when 'large' and 'insect' are combined into 'large insect'. To achieve that, our brain has to retrieve the meaning of each word and combine them in terms of the meaning (semantic composition), but also in terms of structural dependency, i.e., which word is a modifier and which word is modified (syntactic composition).

Different approaches have been taken in cognitive neuroscience to unpack the processes underlying such composition. Some studies looked into the composition of meaningful phrases and sentences as opposed to ones with a semantic or a syntactic violation (for example, lines of work looking at N400, LAN and P600 event-related potential signatures; Kaan, Harris, Gibson, & Holcomb, 2000; Kutas & Federmeier, 2011; Lau, Phillips, & Poeppel, 2008) or as opposed to unstructured word lists (e.g., measuring fMRI BOLD; Friederici, Meyer, & von Cramon, 2000; Hultén, Schoffelen, Uddén, Lam, & Hagoort, 2019; Humphries, Binder, Medler, & Liebenthal, 2006). Other studies manipulated the level of semantic or syntactic complexity of a sentence or a phrase (e.g., Bornkessel, Zysset, Friederici, von Cramon, & Schlesewsky, 2005; Makuuchi, Bahlmann, Anwander, & Friederici, 2009; Pallier, Devauchelle, & Dehaene, 2011). However, processing a longer phrase or sentence necessarily involves some processes in addition to composition such as storage of information in working memory, ambiguity resolution, pragmatic

or discourse inferences etc. One approach that avoids these potential additional processes is looking at composition in minimal two-word phrases ('large insect') as compared to processing a single word ('insect'). In this case, the composition is stripped to the most basic process: as opposed to processing a single word, the only added processes should be retrieval of the meaning of the second word (adjective) and composition. While natural language utterances are clearly more complex, we can take composition of such a minimal phrase as a starting point for investigating brain dynamics supporting linguistic combinatory processing.

In the present study, we investigate semantic and syntactic composition of two-word phrases as compared to processing a single word in Dutch using magnetoencephalography (MEG).

5.1.1 Composition of minimal adjective-noun phrases: MEG studies

A series of MEG studies by Pylkkänen and colleagues investigated spatial and temporal aspects of processing minimal adjective-noun phrases as opposed to processing nouns preceded by consonant strings (e.g., Bemis & Pylkkänen, 2011, 2013a, 2013b; Del Prato & Pylkkänen, 2014; Flick et al., 2018; Pylkkänen, Bemis, & Blanco Elorrieta, 2014; Westerlund & Pylkkänen, 2014; Ziegler & Pylkkänen, 2016 etc.; see Pylkkänen, 2016 for a review). In these studies, participants were presented with either an adjective and a noun (e.g. 'red car' - composition condition) or a meaningless consonant string and a noun (e.g. 'xgf car' - no composition condition; a consonant string equalizes the amount of visual input to the brain in both conditions) in a word-by-word fashion. Subsequently, they were presented with a picture or a written question and judged whether it corresponded to the meaning of the phrase or the noun, depending on the condition. These studies consistently report higher levels of activity in the left anterior temporal lobe (LATL) in the composition condition as opposed to the no composition condition approximately 200-250 ms after noun onset (henceforth, we will refer to this difference as the 'LATL composition effect'). Similar composition-related activity has been reported for the processing of noun-noun compounds (Flick et al., 2018; Zhang & Pylkkänen, 2015) and minimalistic verb phrases (Kim & Pylkkänen, 2019; Westerlund, Kastner, Al Kaabi, & Pylkkänen, 2015). Besides English, involvement of LATL in composition of minimalistic two-word phrases in a similar time-window has also been reported for Modern Standard Arabic (Westerlund et al., 2015) and American Sign Language (Blanco-Elorrieta, Kastner, Emmorey, & Pylkkänen, 2018). Altogether, these results were interpreted as demonstrating that LATL is the brain region most distinctly responsible for composition of the meaning of two words (Pylkkänen, 2016; Pylkkänen & Brennan, 2019). It should be noted, however, that the localization and timing of the LATL composition effect varied considerably between studies (reviewed in *Table 5.5, General*

Discussion). It was sometimes located in areas beyond what would typically be considered the anterior portion of the left temporal lobe. The earliest onset and the latest offset time of the effect varied between roughly 180 ms and 350 ms, lasting for approximately 50-100 ms. The presence of such variability was one of the motivations for conducting the present study. We discuss these points in detail below.

In the above-mentioned series of MEG studies, composition-related brain activity has also been reported in several other regions, in addition to LATL, though not consistently: the right anterior temporal lobe in approximately the same time-window as LATL but also in a later time window (Bemis & Pylkkänen, 2011; Poortman & Pylkkänen, 2016), ventromedial prefrontal cortex at approximately 350-500 ms after noun onset (Bemis & Pylkkänen, 2011; Del Prato & Pylkkänen, 2014) and left angular gyrus (at approximately 350-400 ms after noun onset in the visual modality and 540-590 ms after noun onset in the auditory modality; Bemis & Pylkkänen, 2013a). Given that increased activity in these regions was not observed consistently across studies, these effects have to be considered as less robust than the effect in LATL. To anticipate, for these reasons we will, in the present study, focus primarily on the LATL effect when looking at the contrast between the composition and no composition conditions described above.

5.1.2 Investigating semantic composition of minimal phrases

On a theoretical level, one can distinguish between semantic and syntactic composition processes. The contrast between an adjective-noun phrase and a consonant string-noun combination employed in the above described MEG studies does not allow one to distinguish between these two processes as both are involved in the former and both are absent in the latter. However, because the LATL composition effect appears to be modulated by semantic-conceptual factors (discussed in detail below), it has largely been interpreted as reflecting semantic (or conceptual) composition rather than syntactic composition Pylkkänen, 2016; Pylkkänen & Brennan, 2019; Westerlund & Pylkkänen, 2014).

Besides the MEG studies, several fMRI studies investigated which brain regions show increased levels of activity for semantic composition processes in minimal two-word phrases. In one such study, Price and colleagues (Price, Bonner, Peelle, & Grossman, 2015) presented participants with adjective-noun phrases differing in plausibility (i.e., how meaningful the phrase was - e.g., 'loud car' vs #'moss pony'). The processing of more meaningful phrases was accompanied by a higher BOLD signal in both the left and right angular gyri (BA39[2]) than the processing of less meaningful phrases. This difference was interpreted as evidence for angular gyrus playing a crucial role in semantic composition (see also Price, Peelle,

[2]Here and later we mention the Brodmann Area to which the peak point of the observed difference in BOLD-signal belonged since in the present study analyses we define regions of interest (ROIs) in terms of Brodmann Areas.

Bonner, Grossman, & Hamilton, 2016, for supporting causal evidence from transcranial direct current stimulation). Another study (Matchin, Hammerly, & Lau, 2017) contrasted processing determiner-noun phrases and determiner-pseudoword phrases. Here, the authors reasoned that whereas in the case of determiner-noun phrases semantic composition should take place, such semantic composition would not be possible for the determiner-pseudoword phrases since pseudowords do not carry any associated semantic meaning. More BOLD activity was observed for processing determiner-noun phrases in a small region in the posterior portion of the middle temporal gyrus (BA21), presumably showing that this region is more involved in semantic composition. Yet another study (Schell, Zaccarella, & Friederici, 2017) compared brain activity in adjective-noun phrases (e.g. 'blaues Schiff' [blue ship]), determiner-noun phrases (e.g. 'dieses Schiff' [this ship]), and single nouns (e.g. 'Schiff' [ship]). Schell and colleagues viewed the adjective-noun composition as 'semantically driven' and reasoned that the regions more involved in processing of these phrases as opposed to determiner-noun phrases (which they view as 'syntactically driven' since they have reduced descriptive content) and single nouns should be partaking in semantic composition[3]. Processing the adjective-noun phrases in comparison to single nouns was accompanied by more BOLD activity in the left inferior frontal gyrus (LIFG; BA45), and in comparison to determiner-noun phrases it was accompanied by more BOLD activity in the left angular gyrus (BA39). The regions identified in all of these studies can be seen as additional potential candidates for carrying out semantic composition in minimal adjective-noun phrases: left and right angular gyri, left posterior temporal lobe, and left inferior frontal gyrus.

To investigate semantic composition of two-word phrases further, in this project we will primarily focus on whether we can observe the LATL composition effect, as reported in the MEG studies discussed above. In addition, we will investigate whether the semantic composition effect might be present in the regions indicated by the fMRI studies, though we will do the latter only in an exploratory manner since the regions in which effects were observed in the fMRI studies did not seem to exhibit differential activity in the MEG studies.

Modulation of the LATL composition effect by noun specificity

In an attempt to look further into the specific process that the LATL composition effect reflects, several MEG studies varied the type of stimuli within the experimental set-up outlined above. One factor in this context is noun specificity (West-

[3]Note that Schell and colleagues consider determiner-noun phrase composition 'syntactically-driven' whereas in the design of Matchin and colleagues such determiner-noun phrases are considered to have a semantic component (relative to the determiner-pseudoword condition). Similar to Matchin and colleagues, we think that there would be a fair share of semantic composition for a determiner-noun phrase, but will report the Schell and colleagues study here with the reasoning of the original authors.

erlund & Pylkkänen, 2014; Zhang & Pylkkänen, 2015; Ziegler & Pylkkänen, 2016). In these studies, no composition effect (i.e., no difference between adjective-noun and consonant string-noun combinations) was observed in LATL when the noun in the adjective-noun phrase denoted a more specific meaning (e.g. 'rose', 'trout'), whereas the composition effect *was* present when the noun denoted a less specific meaning (e.g., 'flower', 'fish'; Westerlund & Pylkkänen, 2014; Ziegler & Pylkkänen, 2016). These findings were interpreted as a reflection of a narrowing down of the meaning of the noun: when the noun has a rather specific meaning, the information provided by an adjective provides relatively little additional information concerning potential referents. By contrast, when the noun is less specific, the adjective adds more information with respect to potential referents. The fact that a conceptual feature of the noun appears to modulate the composition effect suggests that the effect reflects semantic composition rather than syntactic composition since phrases with high and low specific nouns do not differ syntactically (Pylkkänen, 2016; Pylkkänen & Brennan, 2019). However, the modulation of the composition effect was observed in somewhat differing regions of interest (within or close to LATL) in the three relevant studies (specifically, left BA38 in Zhang & Pylkkänen, 2015; left BA21 in Ziegler & Pylkkänen, 2016; left BA38+21+20 in Westerlund & Pylkkänen, 2014). Moreover, one of the studies that looked at noun specificity reported finding a confounding difference in adjective-noun plausibility (i.e., typicality) between high and low specificity conditions in a post-hoc norming study that they conducted after their data was collected (Westerlund & Pylkkänen, 2014), whereas the other two do not report controlling for plausibility (Zhang & Pylkkänen, 2015; Ziegler & Pylkkänen, 2016). Thus, though a clear modulation of the LATL composition effect by noun specificity would be important for the interpretation of this effect, it appears that this modulation needs additional investigation to establish its robustness and stability, and concerning potential confounds.

Modulation of the LATL composition effect by adjective class

Differences between types of adjectives might also have a modulating role, and might thus allow for a further specification of the composition processes reflected in the LATL composition effect (Ziegler & Pylkkänen, 2016). The meaning of one class of adjectives, so called *scalar adjectives*, is strongly context dependent - their meaning largely depends on the noun that they are combined with. For example, 'large' refers to different sizes when combined with 'fruit', 'horse' or 'house'. In theoretical semantics, this is captured by the assumption that the noun determines the comparison class for the property denoted by the adjective. For example, in the case of 'large' the comparison class would need to consist of typical sizes of either fruits, horses or houses (e.g., Kennedy, 2007; Kennedy & McNally, 2005; E. Klein, 1980; van Rooij, 2011b). These scalar adjectives are also sometimes called *gradable* or *vague*; other examples are 'tall', 'long', 'loud',

'heavy' etc. By contrast to scalar adjectives, the meaning of so-called *intersective adjectives*[4] does not depend on context, or does so to a much lesser degree. The meaning of an intersective adjective such as 'square', 'dead' or 'ceramic' remains relatively stable for different nouns (naturally, some ambiguity always remains, but we assume that this ambiguity is drastically smaller for intersective compared to scalar adjectives).

Ziegler and Pylkkänen (2016) contrasted adjective-noun combinations comprised of scalar adjectives or intersective adjectives and found more activity in the time-window of the LATL composition effect – 200-300 ms after noun onset – for nouns combined with intersective adjectives than for nouns combined with scalar adjectives (though the ROI was again somewhat different from other studies on the LATL composition effect). Ziegler and Pylkkänen interpret this difference as suggesting that composition of nouns with intersective adjectives happens at the standard time-window of the LATL composition effect, whereas composition of the noun with scalar adjectives does not or cannot yet happen at this time-window, but happens later. Specifically, they propose this interpretation assuming that the activation of semantic features of a noun happens in a gradient fashion, with retrieval proceeding from general to highly specific features (Moss, McCormick, & Tyler, 1997; Pylkkänen, 2016). They further assume that composition of a noun with a scalar adjective requires for the noun meaning to be rather specific since scalar adjectives' meaning depends on a specific comparison class, whereas composition of a noun with intersective adjectives does not require the same degree of specificity because the adjective meaning depends less on the comparison class of the noun. Hence, when participants see a noun preceded by an intersective adjective, composition can happen early, when still relatively little information about the noun is retrieved (at 200-300 ms), whereas when they see a scalar adjective, more information about this noun has to be retrieved before composition can happen, so it happens later (therefore, no composition effect at 200-300 ms yet).

However, the Ziegler and Pylkkänen study is the only one to date that reports the modulation of the LATL composition effect by the adjective class, and also with a region of interest slightly different from the original studies reporting the LATL composition effect. In addition, as in the studies on the role of noun specificity described above, also in this study on adjective types, there was no explicit control for matching the plausibility of the adjective-noun phrases between conditions. Thus, as for the potential effect of noun specificity, it appears that the modulation of the composition effect by adjective class also should be checked for robustness and stability.

[4]In fact, these adjectives are more commonly called 'non-gradable', with 'intersective' being only a subset of such adjectives (Kamp & Partee, 1995; Partee, 1995), but we will follow the terminology used in Ziegler & Pylkkanen article for consistency.

5.1.3 Investigating syntactic composition of minimal phrases

In contrast to semantic composition, syntactic composition has received much less attention in MEG studies, but has been investigated in a number of fMRI studies. Here, we discuss research on syntactic composition with a focus on two-word phrases.

As already mentioned, the experimental design of the MEG studies on the LATL composition effect cannot directly distinguish between semantic and syntactic composition effects. In the respective conditions contrasted in these studies, either both aspects of composition (semantic and syntactic) were simultaneously present or simultaneously absent. However, a number of fMRI studies have shown that LATL is sensitive to manipulations of syntactic information (e.g., Brennan et al., 2012; Brennan & Pylkkänen, 2017; Rogalsky & Hickok, 2009; Schell et al., 2017), and this even holds for sentences consisting exclusively of pseudowords where lexico-semantic information is fully absent (e.g., Friederici et al., 2000; Humphries et al., 2006; Mazoyer et al., 1993; however, see Flick & Pylkkänen, 2018; Pylkkänen & Brennan, 2019 for arguments on why sentences with pseudowords cannot be considered completely void of meaning). Thus, LATL is also a candidate region for performing syntactic composition in minimal phrases.

Several recent fMRI studies focused specifically on identifying brain regions responsible for syntactic composition in minimal phrases. Zaccarella and Friederici (2015) investigated brain activity when processing a two-word phrase consisting of a determiner and a pseudoword in German (e.g., 'DIESE FLIRK' [this flirk]). Syntactically, such a phrase is clearly a determiner phrase, but it has no meaning (i.e., no semantic component). Compared to a two-word list condition (e.g. 'APFEL FLIRK' [apple flirk]) which lacked syntactic structure, the processing of determiner-pseudoword phrases resulted in more BOLD activity in a portion of the left inferior frontal gyrus (BA44). In a similar study that used real words instead of pseudowords, Zaccarella and colleagues (Zaccarella, Meyer, Makuuchi, & Friederici, 2017) observed an increase in activity related to syntactic composition again in LIFG (BA44, BA45), but also in left posterior superior temporal sulcus (BA22). Consistent with these results, the already mentioned study of Schell and colleagues (Schell et al., 2017) reported more BOLD activity for processing determiner phrases in comparison to single word processing and in comparison to adjective-noun phrase processing also in LIFG (BA44, BA45; but also BA47) and in left posterior superior temporal sulcus and middle temporal gyrus (BA21, BA22). Based on these studies, the LIFG as well as parts of the posterior temporal lobe potentially play a crucial role in syntactic composition processes. Note, however, that not all fMRI studies on syntactic composition observed these effects: in another fMRI study that looked at minimal phrase composition, Matchin and colleagues (Matchin et al., 2017) did not observe any differences between processing a two-word phrase and a non-structured list; we will return to this point in the *General Discussion*.

To the best of our knowledge, there is until now no MEG study looking specifically at syntactic composition processes for minimal adjective-noun phrase processing. In the present study, we will therefore include a condition targeting specifically syntactic composition processes in addition to conditions targeting semantic (and syntactic) composition.

5.1.4 Present study

In the present study, we will primarily look into composition processes as reflected by the LATL composition effect reported in previous studies, and, more specifically, into potential modulation of the LATL composition effect by semantic-conceptual properties of the adjective and the noun in adjective-noun phrases, which support its interpretation as reflecting specifically semantic composition. In addition, we will look into syntactic composition processes in adjective-noun phrases. We do so using MEG which will allow us to identify corresponding activity with high temporal resolution and to identify the brain regions that are involved.

The literature review provided above shows that while multiple MEG studies reported a composition effect in LATL (starting from Bemis & Pylkkänen, 2011), these previous findings are mixed in terms of spatial and temporal extent of the effect. In the present study, we will use a design that is parallel to the one used in the MEG studies by Pylkkänen and colleagues. Within this general design, we will compare the processing of adjective-noun phrases with the processing of single nouns (i.e., nouns preceded by a letter string); the contrast between these two conditions should reflect the composition processes in phrases. We will focus on the LATL composition effect, i.e., more activity in LATL at approximately 200-250 ms after noun onset for the adjective-noun phrase as opposed to single nouns. In addition, we will introduce two factors that appear to modulate composition processes and are, thus, important for understanding the observed effect: noun specificity (see Westerlund & Pylkkänen, 2014; Zhang & Pylkkänen, 2015; Ziegler & Pylkkänen, 2016) and adjective class (see Ziegler & Pylkkänen, 2016).

In the present study we chose to closely follow the design and analysis choices of a previous study. Specifically, we closely follow the study by Ziegler and Pylkkanen (2016; henceforth referred to as Z&P) since it included both the noun specificity and the adjective class manipulations that we are interested in. In our analyses, we compared neural activity of processing nouns preceded by real adjectives as opposed to preceded by a letter strings (i.e., single noun processing). We used the same region of interest and time-windows in which the Z&P study reported their observed effects as our pre-specified hypotheses in confirmatory analyses. However, as there is some inconsistency in the localization and timing of the reported composition effect in previous studies, we also planned exploratory analyses of other regions and time-windows where composition effects have previously been reported (see *Method* for details on different analyses that

we conducted).

The second goal of the present study was to investigate syntactic composition processes in adjective-noun phrases within the same set-up as for semantic composition. The involved brain regions and the time-course of syntactic composition with such minimalistic phrases has not yet been addressed in MEG studies. To this end, we included a condition in which instead of an adjective the participants saw pseudowords combined with real nouns (similar to a number of studies that used such 'jabberwocky' stimuli in the past - e.g., Matchin et al., 2017; Mazoyer et al., 1993; Pallier et al., 2011; Zaccarella & Friederici, 2015). These pseudowords were inflected for grammatical gender like real adjectives and are, therefore, henceforth referred to as 'pseudoadjectives'. This was possible because the present study was conducted in Dutch (rather than English, a language where the following described properties are not present). In indefinite noun phrases with a prenominal adjective in Dutch, the adjective agrees with the noun in terms of the noun's grammatical gender - it either has an ending 'e' or does not depending on which gender the noun belongs to (e.g., 'klein paard'[small horse], 'kleine tafel' [small table] but *'kleine paard', *'klein tafel'). The pseudoadjectives in our study agreed with nouns in terms of the grammatical gender, i.e., there was morphosyntactic agreement. We expected that such pseudoadjective-noun phrases would induce syntactic composition in the absence of semantic composition (see *Materials* below for a detailed discussion of a control for potential semantic associations). Given that no previous MEG studies targeted specifically syntactic composition processes, we did not have strong hypotheses about the time-course and regions where they should be observable. We thus compared the neural activity of processing nouns preceded by pseudoadjectives as opposed to nouns preceded by letter strings (i.e., single noun processing) in an exploratory way in the regions that have been implicated for syntactic composition by fMRI studies investigating minimal phrases.

5.2 Method

5.2.1 Participants

This study fell under the ethical approval for standard studies at Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, by 'Commissie Mensgebonden Onderzoek Regio Arnhem-Nijmegen'. The data was collected between February and May 2019. The number of participants, criteria for participation and inclusion in the analyses were set before the start of the data collection. To take part in the experiment, the participants had to be native speakers of Dutch, 18-35 years old, without any language-related impairment (such as dyslexia), had to have normal or corrected-to-normal vision, and should not have taken part in any of the studies where the experimental stimuli were pre-

tested (described below). In addition, the participants had to meet MEG/MRI inclusion criteria at the Donders Institute (specifically, not claustrophobic, no metal in or on the body, no dental wire, no pacemaker etc.). Forty participants were recruited for this study in exchange for a payment according to the local regulations.

To be included in the analysis, participants had to reach overall accuracy on comprehension questions of at least 80% and have at least 25 trials remaining in each experimental condition after artifact rejection (out of 40 in total)[5]. Two participants were excluded from the analyses due to low accuracy; three additional participants were excluded because of too few trials remaining after artifact rejection. The data of one participant was not analyzed because this participant turned out to be over 35 years old. The data of one further participant was lost due to a technical error. In total, 7 participants were excluded from the analyses leaving 33 valid datasets.

The participants included in the analyses were on average 22.2 years old (range 19-28); 10 were male, 23 female; 29 were right- and four were left-handed. We chose to include left-handed participants even though it is not conventional in neuroimaging language research because the likelihood of their language function being right-lateralized is only marginally higher than for right-handed participants (Mazoyer et al., 2014). However, to ensure that we are not observing a specific pattern of results due to inclusion of the left-handed individuals, we additionally ran all analyses in this project excluding the left-handed individuals. In none of the analyses did the results substantially change after excluding the four left-handed participants.

5.2.2 Materials

The present study had a 4 x 2 design with the factors adjective type (scalar, intersective, pseudoadjective, letter string) and noun type (low specificity, high specificity). Examples of experimental materials are provided in *Table 5.1*; all materials and their properties are available for download in the *Supplemental online materials*. The scalar, intersective, and letter string conditions were included as a replication of the Z&P design, whereas the pseudoadjective condition was added in this study in order to investigate syntactic composition with the meaning stripped away. Note that, for easier comparison, in this section and below we provide information about the Z&P set-up in the footnotes whenever we diverged from it.

We selected 40 noun pairs where one noun had a more general meaning - the low-specificity condition, e.g., 'insect' [insect], 'tas' [bag], 'groente' [vegetable] -

[5]Note that to determine whether a participant should be included, we did not look at the number of remaining trials in the conditions in the additional data collection block of the experiment where pseudoadjectives with grammatical gender mismatch were presented (described below).

Table 5.1: Examples of materials in each experimental condition

condition	adjective	low spec noun	high spec noun	components
scalar adjective	klein(e) [small]	insect [insect]	vlinder [butterfly]	sem + syn
intersective adjective	dood(e) [dead]			sem + syn
pseudoadjective	eelm(e)			syn
letter string	gcxp			none

Table 5.2: Noun properties. The value in the brackets indicates standard deviation.

noun type	N common nouns	N neuter nouns	frequency, log10	lexical decision RT	N letters	N morphemes
low-specificity	18	22	2.69 (0.73)	514 (25)	6.75 (2.22)	1.28 (0.51)
high-specificity	22	18	2.63 (0.84)	515 (31)	6.25 (2.28)	1.30 (0.46)

and the other noun had a more specific meaning - the high-specificity condition, e.g., 'vlinder' [butterfly], 'slaapzak' [sleeping bag],'wortel' [carrot]. The high-specificity noun was always in a set-theoretic subset relation to the low-specificity noun according to the categorization in WordNet (PrincetonUniversity, 2010). Low- and high-specificity nouns were matched on log10 frequency (SUBTLEX-NL; Keuleers, Brysbaert, & New, 2010)[6], lexical decision times (Dutch lexicon project; Brysbaert, Stevens, Mandera, & Keuleers, 2016), number of letters and number of morphemes. In Dutch, nouns have a grammatical gender property - they can be either of *common* or *neuter* gender; it was not possible to match the number of low- and high-specificity nouns in grammatical gender exactly, but they were approximately matched. Summary of noun properties is provided in *Table 5.2* [7].

Each of the nouns was presented in all 4 conditions, i.e., in combination with a scalar adjective, an intersective adjective, a pseudoadjective and a letter string. We used 20 unique intersective adjectives, each combined with 2 nouns; 19 unique scalar adjectives of which 17 were combined with 2 nouns and 2 were combined with 3 nouns; 20 unique pseudoadjectives and 20 unique letter strings also combined with 2 nouns each. The scalar and intersective adjectives themselves could not be matched on frequency and length since scalar adjectives are substantially more frequent and shorter than intersective adjectives[8]; see summary of properties in *Table 5.3*.

The combination of each noun with each adjective condition resulted in 320

[6] Z&P did not match nouns in their materials on frequency.

[7] Note that we do not provide results of a test of significance of differences in terms of these properties between conditions as is customary in language research since such a test has limited utility (see Sassenhagen & Alday, 2016).

[8] Z&P also did not match adjectives in their materials on these features.

Table 5.3: Adjective properties. The value in the brackets indicates standard deviation.

adjective type	**N (unique)**	**frequency, log10**	**N letters, common form**	**N letters, neuter form**
scalar	19	3.66 (0.60)	5.11 (0.66)	4.11 (0.66)
intersective	20	2.91 (0.78)	8.40 (2.14)	7.55 (2.01)
pseudoadjective	20	-	5.20 (0.89)	4.15 (0.67)
letter string	20	-	5.10 (0.91)	

experimental trials in the main experiment [9]. The adjective-noun phrases were matched on a number of properties; see summary in *Table 5.4*. For real adjective conditions, initially we generated a list of 259 different adjective-noun combinations, which entered a plausibility pre-test. In this pre-test, we asked participants to give each phrase a score from 1 to 7 on how natural it sounds (more details about methods and participants, data collection and analysis code for each pre-test are available in the *Supplemental online materials*). Every adjective-noun combination received a score from 25 Dutch native speakers recruited in a web-based study. Based on the average scores, we selected 80 scalar adjective-noun phrases and 80 intersective adjective-noun phrases with a matched average plausibility score[10]. In the next step, we wanted to ensure that scalar and intersective adjectives we used are indeed perceived as different classes of adjectives in terms of their context-sensitivity. One possible test of scalarity that was used by Z&P is embedding an adjective in a phrase like 'large for an insect' where a scalar adjective forms a meaningful phrase whereas an intersective one does not (e.g., #'dead for an insect'; Kamp & Partee, 1995; Siegel, 1976), presumably related to the fact that 'for a' requires retrieving a comparison class which is not possible for properties denoted by intersective adjectives (i.e., there are no different levels of

[9] Z&P had 50 nouns in each noun specificity condition, and did not have a pseudoadjective condition, so their experiment had 300 trials in total.

[10] Z&P do not report matching adjective-noun phrases in different conditions on plausibility. To assess whether a potential confound in plausibility of adjective-noun phrases used by Z&P could have been responsible for the observation of different activity levels for different adjectives classes, we also collected plausibility data for the set of materials used in the Z&P study. We collected data for this purpose in the same way as we did for our own materials, only translating the instructions from Dutch into English and recruiting native speakers of US English. We asked participants to give each phrase a score between 1 and 7 on how natural it sounds (more details, data collection and analysis scripts are available in the *Supplemental online materials*). Every adjective-noun combination received a score from 25 participants recruited in a web-based study. The intersective adjective-noun phrases in Z&P materials received a lower mean score - 5.67 (SD 1.03, range 3.08-6.96) - than the scalar adjective-noun phrases - 6.42 (SD 0.48, range 4.44-7; $t(99)=-6.6$, $p<.001$, $d=-0.66$). While adjective-noun phrases that we used also somewhat differ between conditions in terms of plausibility, the difference between them is substantially smaller (intersective adjective condition: mean 5.88 [SD 0.80, range 3.08-7]; scalar adjective condition: mean 6.08 [SD 0.68, range 3.48-6.69]; $t[79]=-1.74$, $p=0.08$, $d=-0.19$).

Table 5.4: Adjective-noun phrase properties. The value in the brackets indicates standard deviation.

adjective type	**noun type**	**transitional probability**	**cosine similarity**	**plausibility**	**scalarity**
scalar	low-spec	0.0009 (0.0021)	0.20 (0.06)	6.02 (0.82)	5.55 (0.74)
	high-spec	0.0005 (0.0010)	0.22 (0.07)	6.15 (0.52)	6.06 (0.57)
intersective	low-spec	0.0015 (0.0037)	0.24 (0.13)	5.88 (0.81)	2.93 (0.76)
	high-spec	0.0011 (0.0024)	0.24 (0.10)	5.89 (0.81)	2.69 (0.78)

being 'dead'). To test this, we followed Z&P and collected data in a pre-test where participants saw each adjective and noun combination in this form: '[adj] for a [noun]', and were asked to judge how natural they sound. We obtained scores for each phrase from 20 different Dutch native speakers in a web-based study. The intersective adjectives received a mean score 2.8, whereas scalar adjectives received a mean score 5.8.[11] The real adjective-noun combinations were also matched on transitional probability from the adjective to the noun, i.e., probability that the adjective is followed by the target noun (calculated based on Dutch Google web n-gram corpus (Brants & Franz, 2009), using a web-interface provided by Information Science department of the University of Groningen). Specifically, transitional probability was defined as adjective-noun bigram frequency divided by the total frequency of the adjective in the corpus minus the cases where the adjective was at the end of the sentence or where it was followed by punctuation. Finally, real adjective-noun combinations were matched between conditions on cosine similarity (similarity scores were computed using R package *LSAfun* (Günther, Dudschig, & Kaup, 2015) based on vectors for each word created using a skipgramm-model trained on Dutch Wikipedia texts taking into account morphology (Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018).

The pseudoadjective condition in our study was conceived as the condition where we should observe morphosyntactic composition processes but no meaning/semantic composition. The ending of the pseudoadjective always agreed with the following noun in terms of grammatical gender marking. The nonwords (pseudoadjectives) were generated based on real words in CELEX database using WordGen software (Duyck, Desmet, Verbeke, & Brysbaert, 2004), making sure that each 2-letter bigram in the nonword actually appears in a real word in Dutch at least 30 times. Afterwards, the ending 'e' was added to these nonwords in cases when they were combined with a noun of common grammatical gender, and no ending when combined with a noun of neuter grammatical gender. In order to discourage semantic composition, we had to be sure that participants do not have any meanings associations with the pseudoadjectives that we used. To test for that, we administered a pre-test in which participants were presented with a

[11] For comparison, Z&P report a mean score 2.54 for intersective adjectives and 5.44 for scalar adjectives.

pseudoadjective and 4 options for nouns that could follow it. The participants' task was to click on the noun that their intuition told them best matched as a continuation of this adjective (the pseudoadjectives were interspersed with real adjectives). One of the 4 possible options was the target noun (i.e., the noun we wanted to use in the experiment) while the other 3 options were randomly selected for every participant and every pseudoadjective from the set of all other gender-compatible target nouns (i.e., nouns that were used as targets for other pseudoadjectives). We reasoned that if a certain pseudoadjective-noun combination has a stable association for participants, they should consistently select the target noun. If there is no stable meaning association, the target noun should be as suitable as the distractor nouns and, hence, be selected at a chance level - around 25% of the time. We obtained judgments from 30 Dutch native speakers for each pair in a web-based study. Because of the low sample size[12], to allow for a margin of error we considered all cases where the target noun was selected by 30% or less of the participants corresponding to a chance level. Since it turned out to be difficult to avoid any pseudoadjective-noun pairs that were selected by more than 30% of participants, we decided to include 4 pseudoadjective-noun pairs (out of 80 in total) which exceeded this threshold. In the selected pseudoadjective-noun pairs, the average proportion of participants who selected the target noun was 21% (range 6-37%).

Another important consideration with pseudoadjective-noun combinations was that since these were nonwords without meaning and there were no function words presented, we could not be sure that participants would indeed carry out the morphosyntactic composition with the nouns for them. In principle, they could also ignore pseudoadjectives and still respond to comprehension questions (see *Procedure* below). As a way to find out whether participants engaged in morphosyntactic composition for pseudoadjective-noun combinations, we added an additional block of 60 trials at the end of the main experiment, i.e., in addition to the main 320 trials. Out of these, in 20 trials the ending of the pseudoadjective mismatched the grammatical gender of the noun (violation condition), in 20 trials the ending of the pseudoadjective matched the grammatical gender of the noun (correct condition) and in 20 filler trials with real adjectives, the adjective-ending agreed with the grammatical gender of the noun. These trials had exactly the same structure as in the main experiment. If participants indeed engage in morphosyntactic processing for the pseudoadjective conditions, we expected to observe an event-related fields (ERF) signature of violation processing when contrasting the violation and correct grammatical gender marking conditions. If participants did not engage in morphosyntactic processing, their processing system should not be sensitive to the mismatch and, therefore, we should not observe a differing ERF pattern. There is no parallel previous study that has looked at

[12]We were restricted by the number of available participants in the pool of Dutch native speakers that we used for pre-tests.

ERF or ERP signatures in response to violations of grammatical gender marking on pseudowords but there is some research that looked at similar effects. Some studies with real words looked at gender agreement violations in adjective-noun phrases (although none in Dutch), and report more negative ERPs for violations approximately 300-500 ms after onset of the mismatching word (the so-called late anterior negativity effect) followed by more positive ERPs for violations approximately 500-800 ms after the onset of the mismatching word (the so-called P600 effect) (Molinaro, Barber, & Carreiras, 2011). Other studies looked at gender agreement violations in determiner-noun phrases in Dutch and also report similar negativity and positivity effects (Hagoort, 2003; although Hagoort & Brown, 1999 report only the positivity effect). Based on these reports, in our study we expected to observe a difference in ERFs for nouns preceded by pseudoadjectives with mismatching gender marking and for nouns preceded by pseudoadjectives with matching gender marking approximately in the 300-500 ms after noun onset time-window. We could not look at later effects since at 600 ms after noun presentation the participants already saw the comprehension question of that trial (since we used the same trial structure for this additional block of trials).

Finally, the letter string condition was intended only as a control with the same amount of visual input but no semantic or syntactic information. For this condition, we simply generated strings of consonants and ensured that they did not form phonotactically legal Dutch words.

5.2.3 Procedure

During the experiment, the participants were seated in a dimly lit magnetically-shielded room, in a chair with the stimuli projected approximately 50 cm away from their eyes. They gave responses using a button-box. MEG data were acquired with a 275-channel whole-brain axial gradiometer system (CTF VSM MedTech) at a sampling rate 1200 Hz with an analog 300 Hz low-pass filter. The position of the participants' head was recorded and monitored by the experimenter in real time throughout the recording, using 3 coils placed on the nasion, and in the left and right external auditory meatus (Stolk, Todorovic, Schoffelen, & Oostenveld, 2013); in case the participant moved their head more than 10 mm from the original position, they were asked to return to it during the next break (see below). Bipolar vertical and horizontal EOG as well as ECG were recorded using Ag/AgCl-electrodes (placed above and below the left eye, to the left of the left eye and to the right of the right eye, on the right shoulder (in supraclavicular fossa) and on the lowest left rib) with an electrode placed on the left mastoid as the ground electrode.

After the MEG recording, an MRI scan for each participant was obtained on one of three 3T MRI scanners (Siemens) available at the Donders Institute using a high-resolution T1-weighted magnetization-prepared rapid gradient-echo pulse sequence. For the MRI acquisition, the participants were wearing the same

in-ear moulds as for the MEG recording, but with a Vitamin E tablet in them; these points in the ear canal along with the nasion were later manually marked and used for co-registration of anatomical data to the head location during MEG recording. The participants' head shape was digitized using a Polhemus Fastrak digitizer, and also used for coregistration.

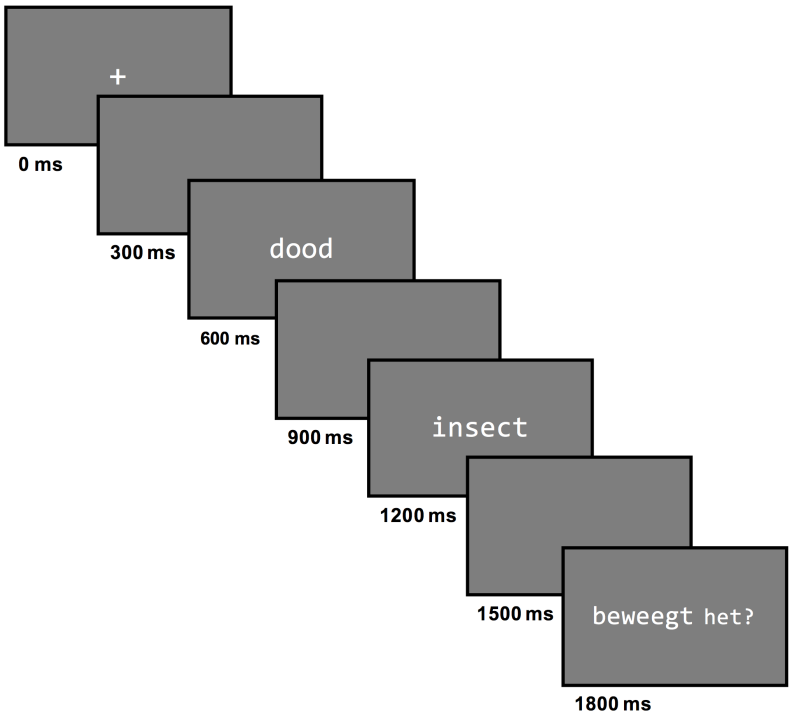
Stimulus presentation was controlled by software package Presentation (Neurobehavioral Systems). The experiment started with the display of instructions, and the participants had a chance to practice on 4 trials. Participants were instructed to 'read the combinations of words and answer questions about them'. They were told that sometimes the first word would be a real word and sometimes not; when the first word would be a real word, the question would be about the meaning of the combination of the two words, whereas when the first word would not be a real word, the question would be about the second word only. The comprehension questions were not full-blown sentences but one or several words followed by a question mark, which the participants were instructed to convert into proper questions themselves (for example, a question like 'has legs?' was supposed to be understood as 'does it have legs?'; we did this in order to avoid participants spending time reading and re-reading long questions and in accordance with what Z&P did in their study). All participants saw all stimuli. The 320 trials of the main experiment were divided into 5 blocks of 64 trials in each; the order of the trials was fully randomized[13]. After these trials, the participants also saw the last block of 60 additional trials some of which contained pseudoadjective violations, with the order randomized within that block. Between blocks, the participants had a chance to rest for as long as they wanted; if needed, during the break the experimenter asked the participants to return their head to the original position.

The trial structure mirrored the one used by Z&P; see *Figure 5.1* for an illustration. Each trial started with a fixation cross displayed for 300 ms. The adjectives and nouns were displayed for 300 ms each with a 300 ms long blank before, between and after. The adjective and noun were displayed in the center of the screen, in white 36 pt *Consolas* font on gray background. The comprehension questions remained on the screen until the participant gave a response by pressing one of the two buttons - 'yes' or 'no'. The expected answer to 50% of the questions was "yes" and to the other 50% it was "no". The participants were instructed to avoid blinking during the display of the words and to try to only blink during the question display.

The MEG recording itself took approximately 30 minutes. Each session, including preparation, practice, MEG recording and MRI scanning took approxi-

[13]Note that in Z&P the blocks were constructed manually such that no two adjectives or nouns were the same within each block, then the order of stimuli were randomized within blocks and the order of blocks was randomized; we preferred to instead fully randomize the order of stimuli

Figure 5.1: Trial structure.



mately 1 hour and 40 minutes.

5.2.4 Data pre-processing

MEG data pre-processing

The MEG data was processed using FieldTrip toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011) in MATLAB 2018b environment; all analysis scripts, raw and processed data are available for download in the *Supplemental online materials*. For pre-processing and source reconstruction, we followed the steps described in Z&P as closely as possible, deviating only in adding steps that we considered important to get a meaningful outcome even though they were not mentioned in the Z&P paper (again, whenever we deviated from their analysis steps, it is explicitly mentioned in the text below). The continuous MEG data was split into epochs of 700 ms before noun onset and 600 ms after noun onset (hence, containing both adjective and noun presentation windows). The epochs containing SQUID jump artifacts were excluded. A 1 Hz high-pass filter was applied to the data. All epochs where the signal exceeded 3000 fT were excluded. Epochs containing eye blinks were excluded using a manual and semi-automatic artifact detection procedure. Subsequently, we used ICA to determine and exclude the components corresponding to the heartbeat signal in the data. Finally, we inspected each trial

in a semi-automatic fashion and excluded trials with remaining large muscle or other artifacts[14].

Anatomical data pre-processing

In order to achieve better spatial localization from MEG source reconstructions, we used individual T1-weighted MRI scans of participants to create a volume conduction model of the head and a cortical sheet based source model[15]. The volume conduction model was created as a realistically shaped single shell approximation of the inside of the participants' skull (Nolte, 2003), using FieldTrip toolbox.

As the source model, we constructed a triangulated cortical mesh using FreeSurfer's automatic surface extraction pipeline *recon-all*[16]. The resulting high resolution meshes were surface-registered to a common template and downsampled to a resolution of 7842 vertices per hemisphere, using HCP workbench (Marcus et al., 2011). This procedure resulted in participant-specific source models, in which individual dipole locations could be directly compared across participants.

Source reconstruction

For source-level activity reconstruction, data pre-processed as described above was low-pass filtered at 40 Hz, downsampled to 1000 Hz, and baseline-corrected for 100 ms before critical word onset (i.e., either before adjective onset for the analyses of adjective time-window or before noun onset for the analyses of the noun time-window)[17]. Epochs were then averaged across trials in each condition separately. The forward solution was computed using the individual source and volume conduction models and the participant-specific gradiometer positions (Nolte, 2003). Source activity was estimated for averaged data, in each condition separately, using L2 minimum-norm estimates (Dale et al., 2000; Hämäläinen & Ilmoniemi, 1994). This step employed a prewhitened noise covariance matrix, estimated from the 100 ms preceding the critical word onset using data across all conditions. The result was dipole activity estimates for 7842 positions per hemisphere. In order to be able to investigate activity in specific Brodmann Areas (BA), as was done by Z&P, the cortical source models were parcellated into 374 areas using an adjusted version of the Conte69 atlas[18] (Glasser & Van Essen, 2011; Van Essen, Glasser, Dierker, Harwell, & Coalson, 2012), in which each of

[14]Z&P do not report dealing with heartbeat signal or muscle artifacts, but only report removing trials that were contaminated by eye movement artifacts.

[15]Z&P used a template brain for source reconstruction.

[16]See *http://surfer.nmr.mgh.harvard.edu/fswiki/ReconAllTableStableV6.0* for details.

[17]Baseline correction is not mentioned in Z&P.

[18]Note that Z&P used the Talairach atlas for separating activity into BA regions. Given that both atlases map Brodmann Areas, we do not expect that using a different atlas for parcellation will make a substantial difference in the observed results.

the larger Brodmann Areas was subdivided into slightly smaller parcels; the activity of points falling within each BA region was averaged to produce a single time course of activity in that region.

Sensor-level data

For the sensor-level analyses, the data pre-processed as described above was low-pass filtered at 40 Hz, downsampled to 1000 Hz, and baseline-corrected for 100 ms before critical word onset (i.e., either before adjective onset for the analyses of adjective time-window or before noun onset for the analyses of the noun time-window). The trials from each condition were averaged for each of the participants. We computed synthetic planar gradients from axial gradiometer data, and combined the vertical and horizontal components of the planar gradients for easier interpretation of the group results (Bastiaansen & Knösche, 2000).

5.2.5 Data analysis: Semantic composition effect and its modulation by noun specificity and adjective class

To investigate the semantic composition effect, we performed confirmatory analyses that were directly based on the findings reported by Z&P, as well as exploratory analyses, part of which were based on our expectations from the literature and part of which was unconstrained. Whereas we will be able to interpret the results from the confirmatory analyses with confidence, all results observed with exploratory analyses will need to be verified in future research to be convincing.

Confirmatory analyses

For confirmatory analyses, we assumed that if the effect reported by Z&P is present in the parallel set-up in Dutch, we should be able to observe it in the same region of interest and time-window. For each participant we extracted the activity in left BA21, the ROI used in Z&P, averaging across vertices and time points of the cluster with the largest test statistic that they identified using cluster-based permutation analysis. We then conducted analyses with the extracted activity value as the dependent variable, expecting to observe a significant difference in the same direction as reported by Z&P (analyses performed using *R* (R Core Team, 2018) and the *ez* package (Lawrence, 2016).

Specifically, for the LATL composition effect (i.e., difference in processing a noun preceded by a real adjective and a noun preceded by a letter string) Z&P conducted separate analyses for each adjective and noun type combination as compared to the corresponding letter string condition. They reported more activity at the noun in the condition where an intersective adjective was combined with a low specificity noun as compared to the condition where a letter string was

combined with a low specificity noun; the cluster with the largest test statistic was in the time-window 200-258 ms after noun onset. They did not observe any effects when contrasting other types of adjectives and nouns with the corresponding letter string condition. We extracted the left BA21 activity values for the time-window 200-258 ms after noun onset in each of the conditions and performed four t-tests in parallel to their analyses, one for each of the adjective class and noun combination as opposed to the corresponding letter string and noun condition.

The second effect that we expected was a difference between adjective classes and noun types. In a 2 (scalar vs. intersective adjective) X 2 (low vs. high specificity noun) analysis, Z&P reported more activity in left BA21 for nouns preceded by intersective adjectives as opposed to nouns preceded by scalar adjectives (i.e., main effect of adjective class); with the cluster with the largest test statistic at 200-317 ms after noun onset. In addition, they reported an interaction between adjective class and noun specificity whereby noun specificity modulated left BA21 activity when nouns were preceded by scalar adjectives, but not when preceded by intersective adjectives; here the cluster with the largest test statistic was at 350-471 ms. We extracted left BA21 activity in each of these two time-windows. We ran a parallel 2x2 design repeated measures ANOVA analysis for each of the time-windows.

Exploratory analyses

In the confirmatory analyses, we strictly adhered to the ROI and time-window reported in Z&P study. It is, however, possible that the effect has a somewhat different timing in Dutch or that, given the variability in ROIs used in previous studies, the particular ROI that was used by Z&P does not capture the effect well. For this reason, in addition to the confirmatory analyses, we performed a series of exploratory analyses. For these exploratory analyses, we collapsed trials across different adjective classes and noun types, thus comparing simply neural activity at any noun when it was preceded by any real adjective as opposed to when it was preceded by a letter string. In case we did observe a composition effect for this contrast, we planned to investigate it further by dividing the activity level in the observed region and time-window into separate adjective and noun levels. This latter step would be used to determine whether the observed composition effect is modulated by semantic-conceptual factors, i.e., reflects semantic processing.

In the first step, we inspected a wider range of possibilities in both time and space (ROI) but still with certain constraints given what we know from the literature. Based on our review of previous studies, we chose four ROIs and performed the analysis for each of these regions separately: left BA21 that was used as an ROI in Z&P, BA38, a region which was analyzed as the ROIs in many of the other MEG studies (e.g., Blanco-Elorrieta & Pylkkänen, 2016; Del Prato & Pylkkänen, 2014; Pylkkänen et al., 2014; Zhang & Pylkkänen, 2015), left inferior frontal gyrus - specifically BA44 and BA45 - that was reported as

the locus of the composition effect in an fMRI study (Schell et al., 2017), and, finally, left angular gyrus (BA39) which was also implicated by fMRI studies (Price et al., 2015; Schell et al., 2017). For activity in each of these regions, we ran a paired t-test based cluster-based permutation analysis (Maris & Oostenveld, 2007), looking for significant differences in the time-window 100-500 ms after noun onset[19] to capture a wider range of possible time-windows. We still expected the composition effect to manifest itself as more activity in case a real adjective is combined with a noun than in case a letter string is combined with a noun; thus, our tests were one-sided. In these analyses, consecutive time-points were grouped into clusters when they showed an effect at an uncorrected level $p<0.3$[20]. A corrected, Monte-Carlo significance probability was calculated as the proportion of 10 000 random permutations yielding a cluster with a higher test statistic than the cluster with the highest test statistic in the actual data. Given that we ran four separate tests, we Bonferroni corrected the significance level to be $0.05/4 = 0.0125$. Note that because in these analyses we collapsed data across different levels of adjective class and noun specificity, we had an unequal number of trials in the two conditions (we had 160 trials with real adjective-noun phrases but only 80 trials with letter string-noun phrases). In order to equalize signal-to-noise ratio in the data, we performed source reconstruction for the real adjective-noun phrase condition using a randomly selected subsample of real adjective trials for each participant, with the number of trials here being the same as the number of trials available in the letter string-noun phrase condition for the same participant. To ensure that we are not missing anything because of a specific selected subsample of real adjective-noun phrase trials, we ran 100 different iterations of the analysis, each time selecting a new random subsample of real adjective-noun phrase trials for source activity reconstruction. If an effect is present in at least 80% of the analysis iterations, we would consider this effect reliable.

In the second step, we conducted an unconstrained exploratory analysis looking for potential composition effects in the whole brain at the level of individual dipole sources. Here, we conducted cluster-based permutation analysis looking for both spatial and temporal clusters in the time-window 100-500 ms after noun onset. This test was also one-sided, looking for more activity in the condition where a real adjective is combined with a noun in comparison to a condition where a letter string is combined with a noun. Cortical locations and time-points that showed an effect at an uncorrected level $p<0.05$ were grouped into clusters. A corrected p-value was calculated based on 1000 permutations (the number of

[19]We constrained our time-windows to at least 100 ms after word onset since based on what we know about language processing, we do not expect to observe any effects at an earlier point; constraining the time-window this way allowed our analyses to be more focused and, therefore, more sensitive.

[20]This is the criterion used in Z&P and other studies by Pylkkänen and colleagues. Note that we in addition ran analyses with a more common criterion $p<0.05$, and there were no substantial differences in the results.

permutations was lower than in other analyses presented here since this analysis is significantly more time-consuming than the spatially constrained ones). We considered p-values below 0.05 as significant.

5.2.6 Data analysis: Syntactic composition effects

To investigate brain activity related to syntactic composition, we again performed an analysis constrained by what we know from literature, as well as a whole brain analysis. Prior to exploring the syntactic composition effects, we analysed the ERFs at sensor level at the nouns in the last block of the experiment where pseuadoadjectives with matching and mismatching grammatical gender marking were presented. If we observe an agreement violation effect when comparing matching and mismatching trials, we would be convinced that our participants performed syntactic composition in the pseudoadjective condition. In absence of such violation effects, we cannot be sure that syntactic composition indeed happened in this condition as we intended, so any effects that we observe in the analysis intended to look at syntactic composition cannot be confidently interpreted as reflecting necessarily syntactic composition.

Morphosyntactic agreement violation

The ERFs were calculated at the noun presentation time-window of trials in the last block, for the condition where the noun was preceded by a pseudoadjective with a matching grammatical gender and for the condition where the noun was preceded by a pseudoadjective with a mismatching grammatical gender. We performed a paired t-test based cluster-based permutation analysis comparing ERFs, looking for spatiotemporal clusters between 100-600 ms after noun onset. For cluster selection, time-points and sensors showing an effect at an uncorrected level $p<0.05$ were grouped into clusters; minimum two neighbouring channels showing an effect were required to form a cluster. A corrected significance probability was calculated based on 5000 random permutations.

Exploratory analyses

We compared source-reconstructed neural activity for processing a noun preceded by a pseudoadjective as opposed to when preceded by a letter string. Based on our literature review, we identified three candidate regions in which we expected to observe a syntactic composition effect: left inferior frontal gyrus (BA44, BA45), left anterior/middle temporal lobe (BA21) and left posterior temporal lobe (BA22). For each of these regions separately, we ran an analysis with the same set-up as the one described for semantic composition effects above, expecting higher activity levels for the nouns preceded by pseudoadjectives as compared to the nouns preceded by letter strings. Because the number of trials in each condition was

approximately equal, there was no need to subsample trials and run multiple iterations of this analysis. We looked for temporal clusters between 100 and 600 ms after noun onset. Given that we conducted a separate test for each of the regions, the significance criterion level was Bonferroni corrected to $0.05/3 = 0.016$.

We subsequently ran an analysis looking for potential syntactic composition effect in the whole brain, at the level of individual dipole sources. We expected to observe more activity for nouns preceded by pseudoadjectives as compared to nouns preceded by letter strings. We conducted analysis looking for spatial and temporal clusters in the time-window 100-600 ms after noun onset, with the same set-up as for the whole brain analysis for the semantic composition effect.

5.3 Results

5.3.1 Behavioral results

Participants included in the analysis gave an expected response to the comprehension questions on average on 91% of trials (SD 3.4; range 83%-96%). The mean response time was 1.56 seconds (SD 0.56; range 0.8-3.3 seconds). The purpose of the comprehension questions was only to make sure that participants pay attention to the task, and, therefore, no further analyses on behavioral data were conducted.

5.3.2 Trial exclusion

The artifact rejection steps resulted in exclusion of 10% of trials overall (range 2-24%)[21].

5.3.3 Semantic composition effect and its modulation by noun specificity and adjective class

Confirmatory analyses

For the LATL composition effect with different adjective class and noun specificity combinations at the time-window 200-258 ms after noun onset, we did not observe a significant difference in any of the comparisons: intersective adjectives vs. letter strings combined with low specificity nouns (t[32]=-1.05, p=0.3), intersective adjectives vs. letter strings combined with high specificity nouns (t[32]=0.07, p=0.9), scalar adjectives vs. letter strings combined with low specificity nouns (t[32]=1.07, p=0.3), scalar adjectives vs. letter string combined with a high specificity nouns (t[32]=0.9, p=0.4). The plots depicting the levels of activity in left BA21 for each of these comparisons are provided in Figure 5.2. We thus did

[21]For comparison, Z&P removed 24% of trials overall.

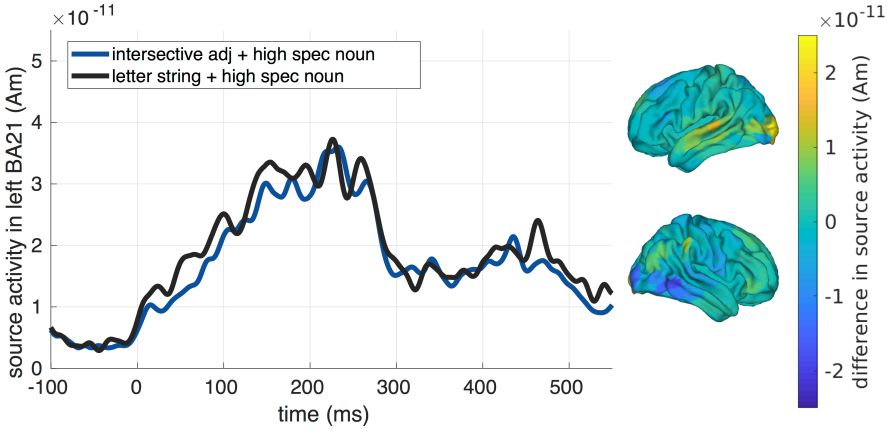
not find evidence for an LATL composition effect or its modulation by adjective class and noun specificity in the region and time-window where such an effect was reported by Z&P.

We did not observe any differences between adjectives classes and noun specificities in the time-window 200-317 ms after noun onset where we expected the main effect of adjective class (main effect of adjective class: F[1,32]=0.2, p=0.6; main effect of noun specificity: F[1,32]=3.1, p=0.08; interaction: F[1,32]=1.2, p=0.2) or in the time-window 350-471 ms after noun onset where we expected to see an interaction effect (main effect of adjective class: F[1,32]=2.9, p=0.09; main effect of noun specificity: F[1,32]=2.1, p=0.1; interaction: F[1,32]=0.2, p=0.6). Figure 5.3 depicts source-reconstructed activity for different levels of adjective class and noun specificity. We thus did not find evidence for differences between adjective classes and noun specificity in the region and time-windows where such effects were reported by Z&P.

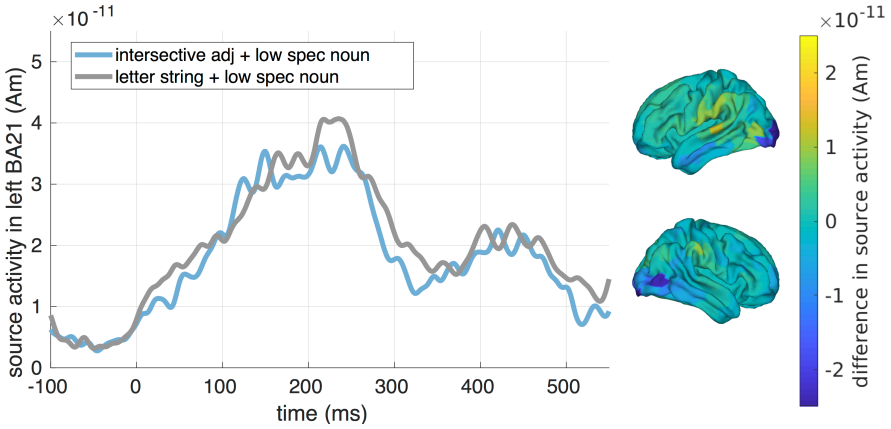
Exploratory analyses

In the constrained exploratory analysis, we did not observe any difference in the source reconstructed activity between a noun preceded by a real adjective and a noun preceded by letter strings in any of the regions of interest that we selected, in none of the iterations with different subsamples of real adjective trials (100 iterations). Neither did we observe any significant differences in the unconstrained exploratory analysis over the whole brain. Figure 5.4 depicts the activity in the whole brain for consecutive time-windows within the window of interest.

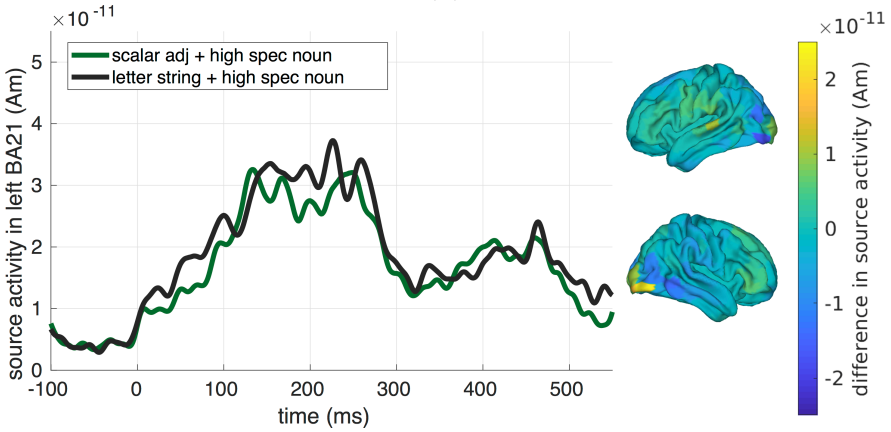
Figure 5.2: Source reconstructed activity levels in the left BA21 for conditions where the noun was preceded by a real adjective and by a letter string. Note that the time-point 0 here corresponds to the onset of the noun. The whole brain plots show activation in the corresponding real adjective condition minus activation in the letter string condition averaged between 200-250 ms after noun onset.



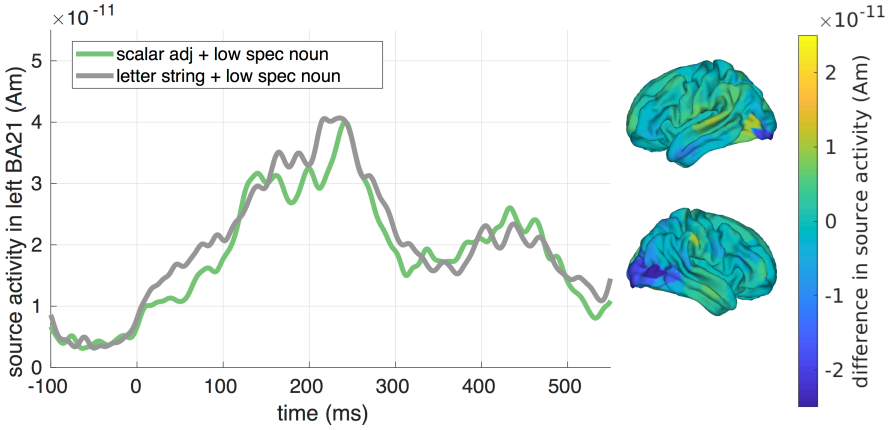
(a)


×10^-11
intersective adj + low spec noun
letter string + low spec noun
source activity in left BA21 (Am)
time (ms)
×10^-11
difference in source activity (Am)

(b)

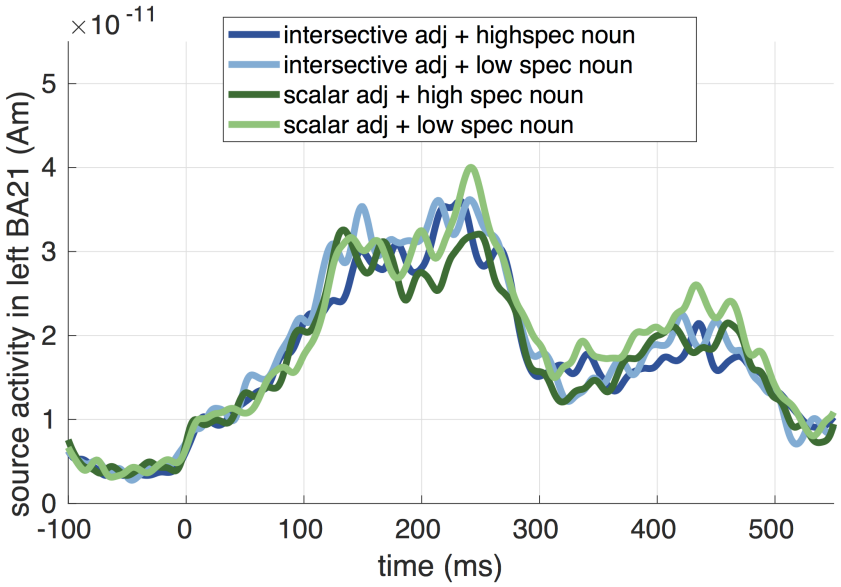

×10^-11
scalar adj + high spec noun
letter string + high spec noun
source activity in left BA21 (Am)
time (ms)
×10^-11
difference in source activity (Am)

(c)

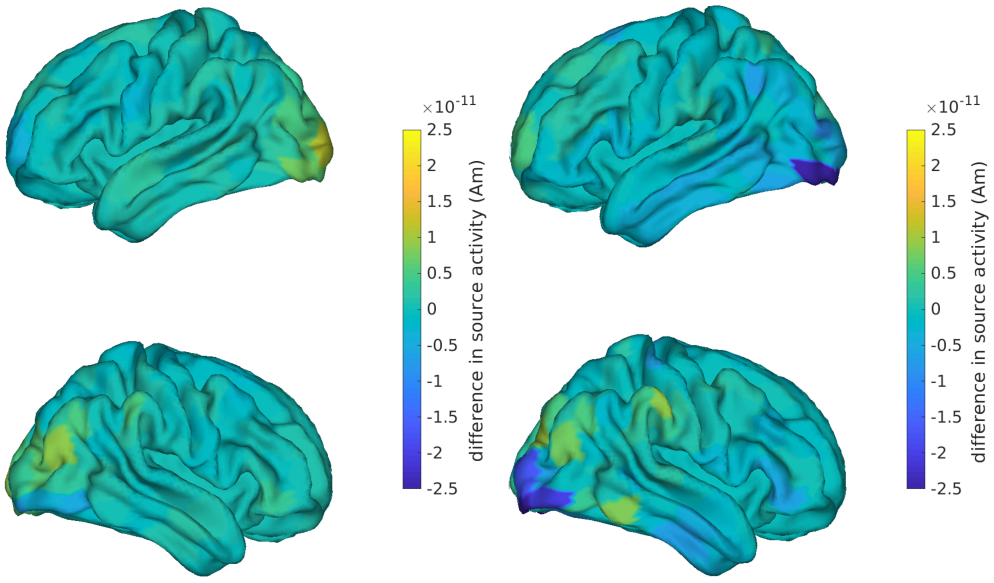

×10^-11
scalar adj + low spec noun
letter string + low spec noun
source activity in left BA21 (Am)
time (ms)
×10^-11
difference in source activity (Am)

(d)

Figure 5.3: Source reconstructed activity levels for different adjective class and noun specificity conditions. Note that the time-point 0 here corresponds to the onset of the noun.



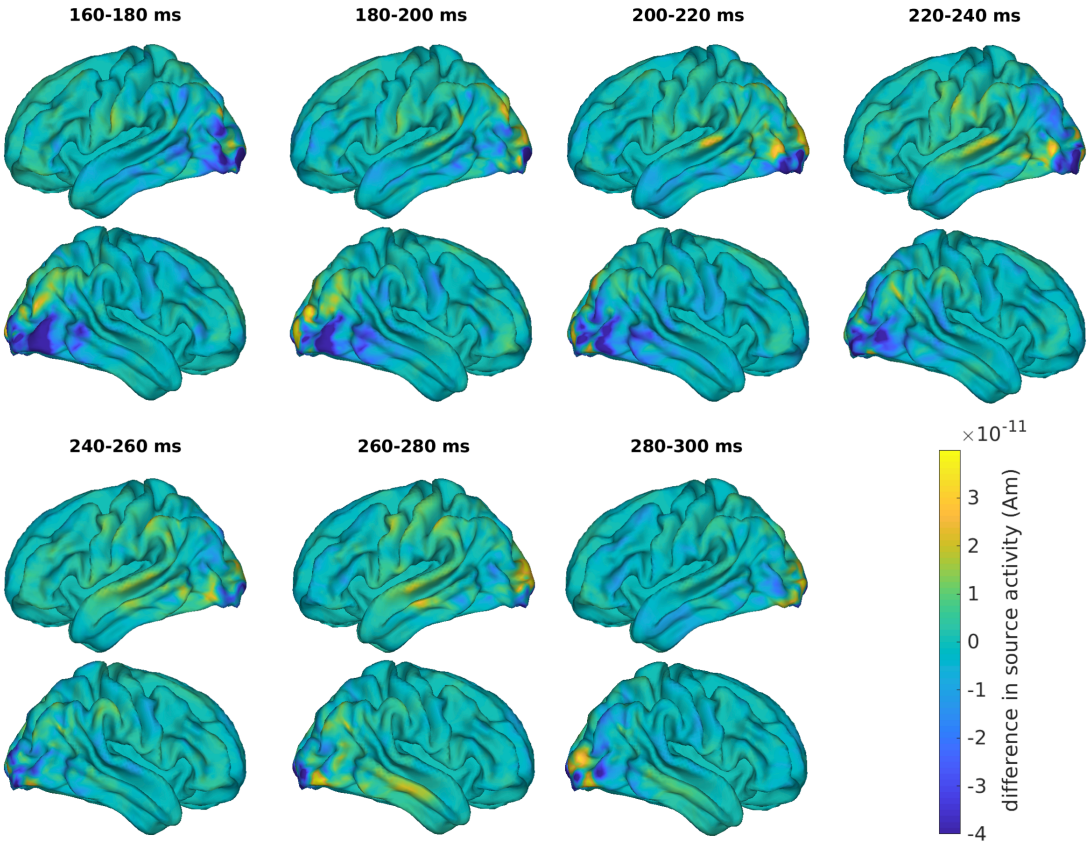
(a) Levels of activity in the left BA21 for different adjective class and noun specificity conditions.



(b) Whole brain plot for: activity for nouns preceded by intersective adjectives minus nouns preceded by scalar adjectives, in the time-window 200-317 ms after noun onset, averaged across noun specificity.

(c) Whole brain plot for: activity for high specificity nouns preceded by scalar adjectives minus low specificity nouns preceded by scalar adjectives, in the time-window 350-471 ms after noun onset.

Figure 5.4: Source-reconstructed activity for nouns preceded by a real adjective minus nouns preceded by a letter string. Note that the time-point 0 here corresponds to the onset of the noun. Note that for this plot we used source-reconstructed activity from only one iteration of the analysis (i.e., one subsample of trials of the intersective and scalar adjective conditions.)



5.3.4 Syntactic composition effect

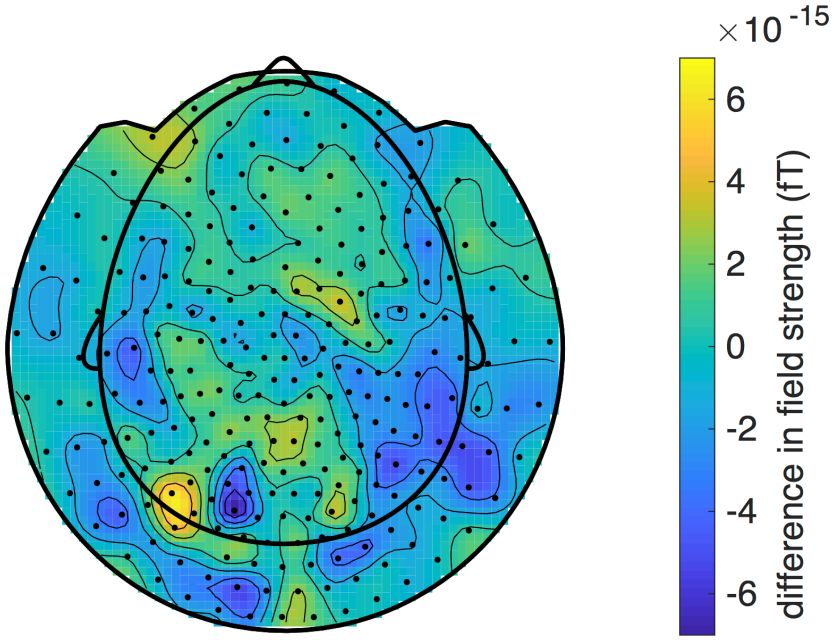
Morphosyntactic agreement violation

We did not observe a significant difference between ERFs at the nouns preceded by a pseudoadjective with matching grammatical gender ending and those at the nouns preceded by a pseudoadjective with mismatching grammatical gender. The ERFs for each of the conditions are depicted in Figure 5.5. Given this result, there is no clear indication that participants noticed a mismatching agreement for the pseudoadjective-noun pairs, and we, therefore, cannot be sure that our participants performed syntactic composition in case of the pseudoadjective condition in our main experimental trials as intended. In contrast, note that we are certain that participants carried out adjective-noun composition in the conditions with the real adjectives given their behavioral performance with the comprehension questions.

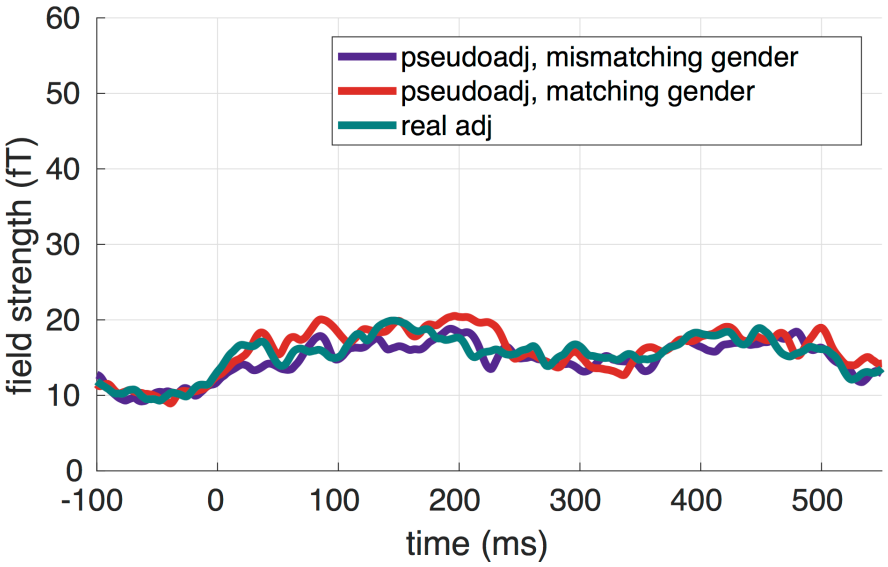
Exploratory analyses

We did not observe any difference in the source reconstructed activity between a noun preceded by a pseudoadjective and a noun preceded by letter strings in any of the regions of interest that we selected. Plots depicting activity levels in each of our regions of interest are provided in Figure 5.6. Neither did we observe any significant differences in the unconstrained exploratory analysis over the whole brain. Figure 5.7 depicts the activity in the whole brain for consecutive time-windows within the window of interest.

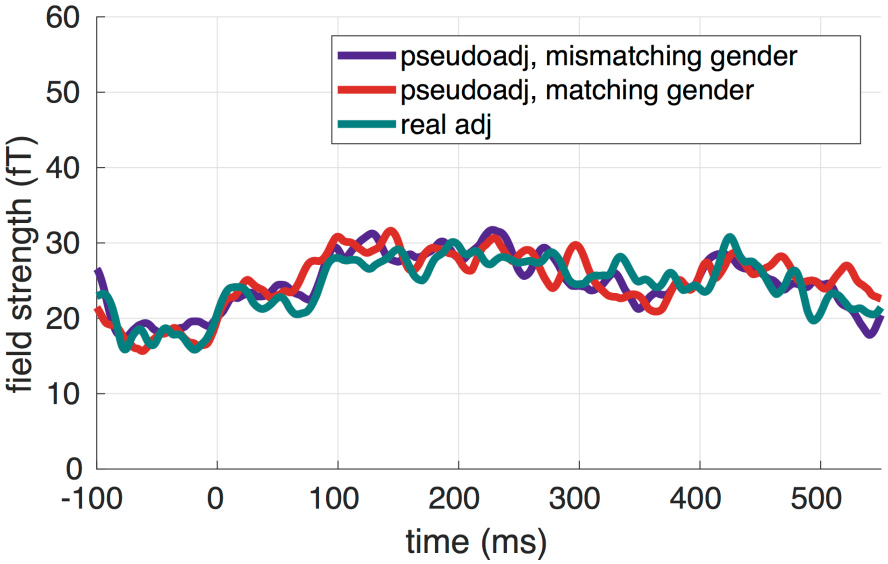
Figure 5.5: ERFs at the nouns preceded by matching pseudoadjective, mismatching pseudoadjective and real adjective (always matching) trials presented in the last, additional block of the experiment. Note that ERFs in the real adjective condition are depicted in plots B and C, but were not included in the analysis. Note that the time-point 0 here corresponds to the onset of the noun.



(a) Distribution of mean amplitude of grand average ERFs at 300-500 ms after noun onset in the matching pseudoadjective condition minus in the mismatching pseudoadjective condition

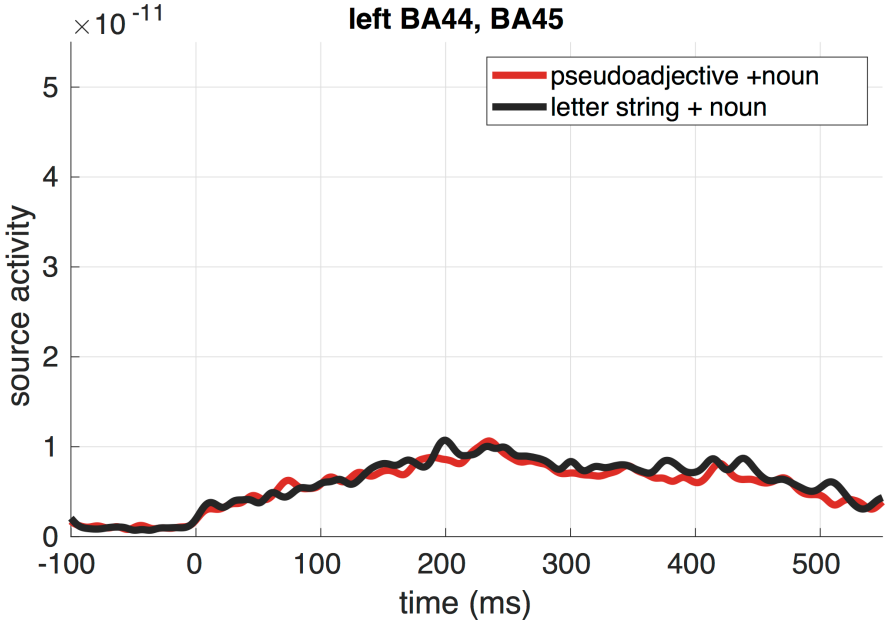


(b) Grand average ERFs at the frontal midline sensors.

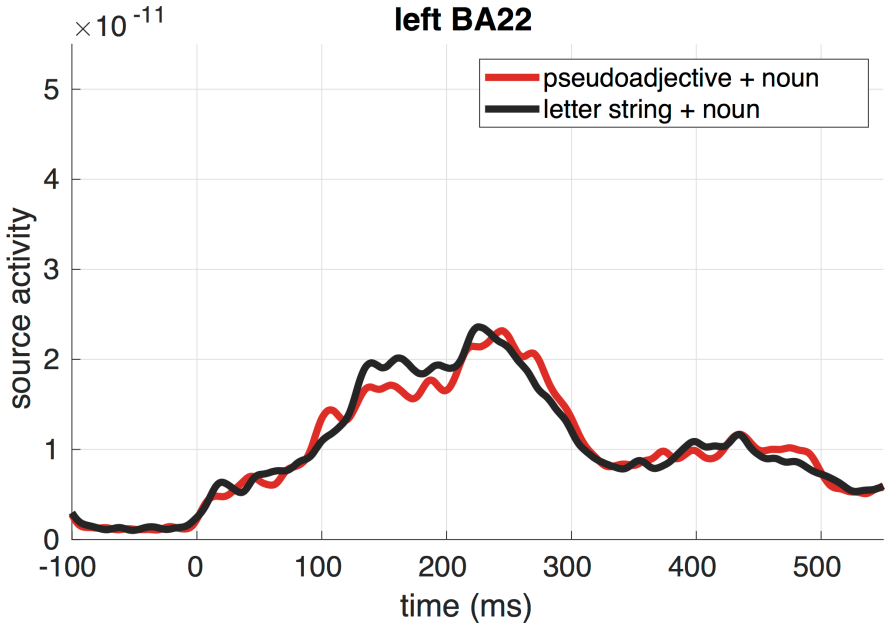


(c) Grand average ERFs at the central midline sensors.

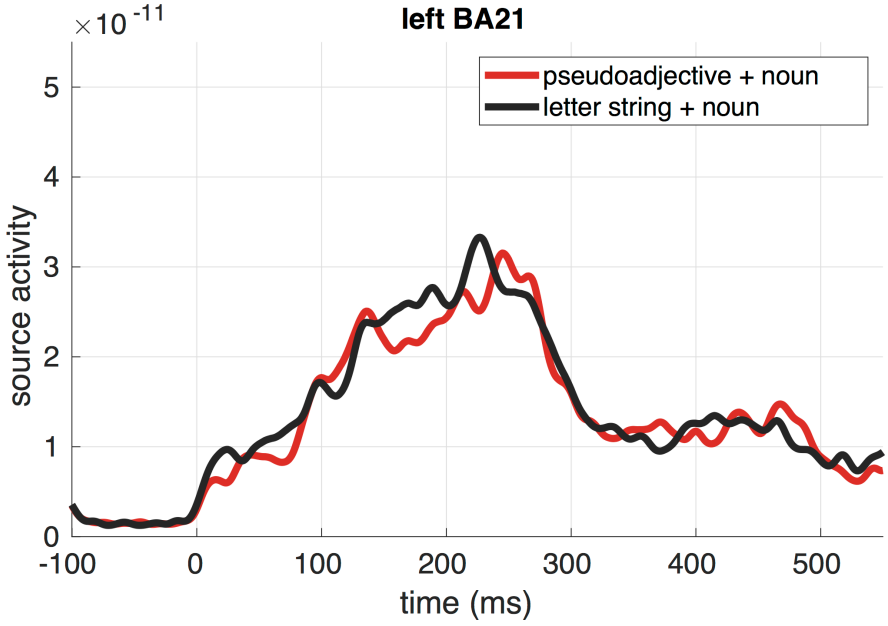
Figure 5.6: Source-reconstructed activity levels in the pseudoadjective and letter string conditions, in the regions of interest chosen for analyses of syntactic composition effects. Note that the time-point 0 here corresponds to the onset of the noun.



(a) Mean activity in left BA44 and BA45.

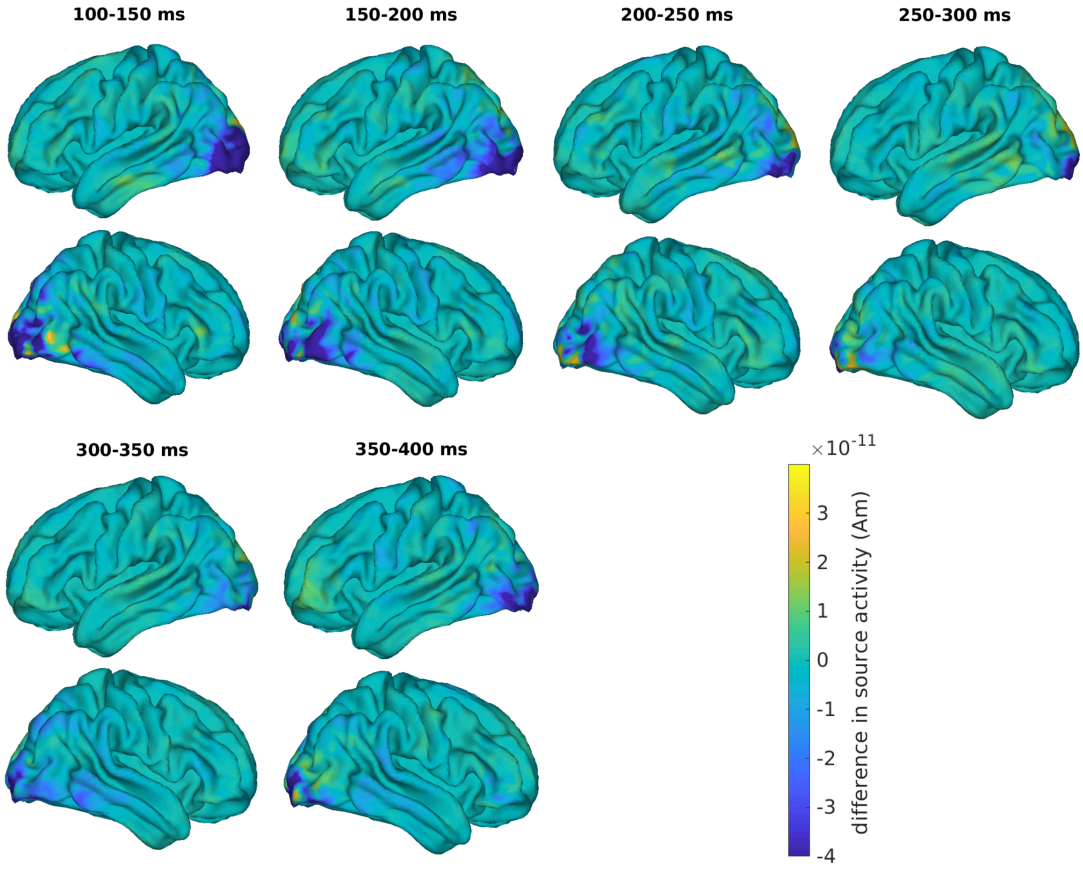


(b) Mean activity in left BA22.



(c) Mean activity in left BA21.

Figure 5.7: Source-reconstructed activity for nouns preceded by pseudoadjectives minus nouns preceded letter strings. Note that the time-point 0 here corresponds to the onset of the noun.



5.3.5 Validation of data processing pipeline

Before discussing our results and their implications, we take an additional step to check the validity and soundness of our data processing pipeline and analysis procedures. To do so, we looked for independent effects in our data, based on a comparison between real adjectives and pseudoadjectives. We call these 'sanity check' analyses.

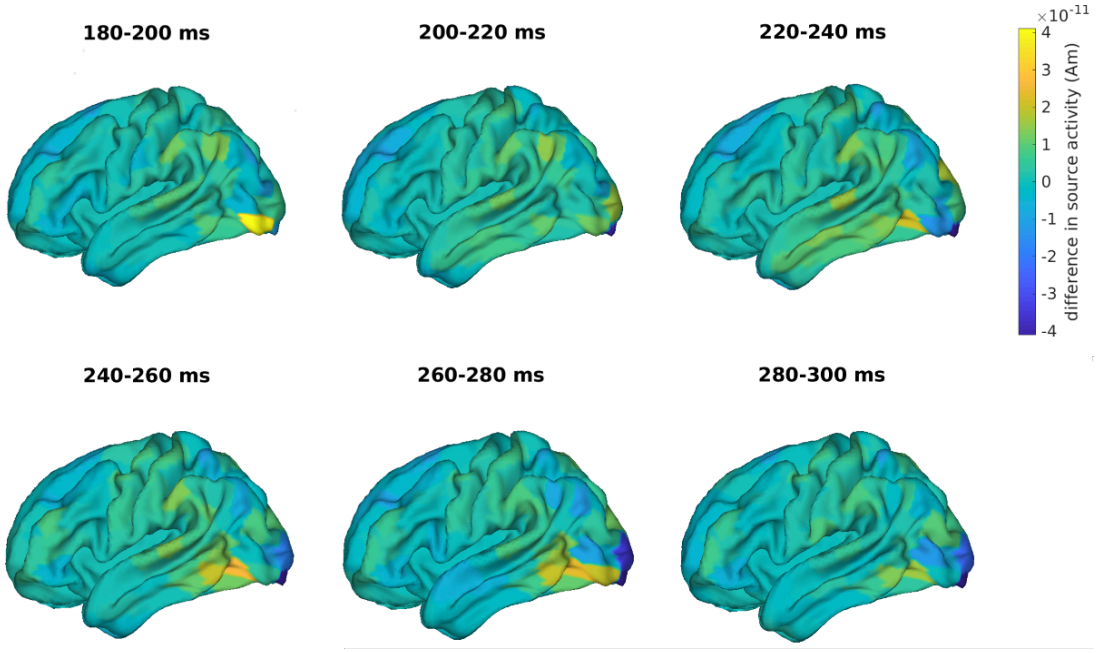
We compared the brain response to the presentation of real adjectives and the response to the presentation of pseudoadjectives or letter strings, at a time-point before the participants saw the noun. At this point, the participants have only seen one word, so no composition effects are expected. Given their task, what participants did at this point is comparable to completing a lexical decision task: they attempted to retrieve lexical information about the word from memory, i.e., recognize the presented word. Hence, we believe that at this point we should observe effects similar to the ones that are reported for the lexical decision task where typically participants respond 'yes' or 'no' to whether a presented word is a real word.

For source reconstruction procedure

An MEG study by Hauk and colleagues (Hauk, Coutout, Holden, & Chen, 2012) employed a contrast that comes close to ours, and is similar in terms of the source reconstruction procedure. In their study (Experiment 2), participants were deciding whether a presented word was a real word and gave responses using eye blinks. Important differences from our study were the following: this was a pure lexical decision task (i.e., nothing followed after participants decided whether the presented word was real), stimuli were presented for just 100 ms, and the real words that participants saw were nouns. For the source-reconstructed data, Hauk and colleagues report significantly higher neural activity when processing real words as opposed to pseudowords in the left middle and inferior temporal lobes between 180-220 ms after word onset.

As a parallel contrast to this one, in our sanity check analysis we compared processing of scalar adjectives and pseudoadjectives at the time-window just following adjective onset. We did not look at intersective adjectives in this analysis because they were on average longer than pseudoadjectives (8.4 letters as opposed to length 5.1 and 5.2 letters in case of scalar adjectives and pseudoadjectives respectively) since we know that word length can influence early word processing (e.g., Assadollahi & Pulvermüller, 2003; Hauk, Davis, & Pulvermüller, 2008; Hauk & Pulvermüller, 2004; Schurz et al., 2010). Because we performed the analyses at Brodmann Area level parcels, based on visual inspection of the effect observed by Hauk and colleagues (Hauk et al., 2012, Figure 6) we estimated that we should see different activity for the two conditions in the left BA37 and BA22 in our parcellation. We ran a paired t-test based cluster-based permutation analysis in

Figure 5.8: Source-reconstructed neural activity for scalar adjectives minus pseudoadjectives. Note that the time-point 0 here corresponds to the onset of the adjective.



each of these regions, looking for temporal clusters between 100 and 300 ms. Note that the data for this analysis was baseline-corrected using a window of 100 ms before adjective onset. We ran one-tailed tests because we specifically expected to see more activity in BA37 and BA22 for scalar adjectives. The set-up of this analysis was parallel to the one used for semantic and syntactic composition effect analyses at the source-level. Given that we ran the test separately on two regions, we Bonferroni corrected the significance level to be $0.05/2 = 0.025$.

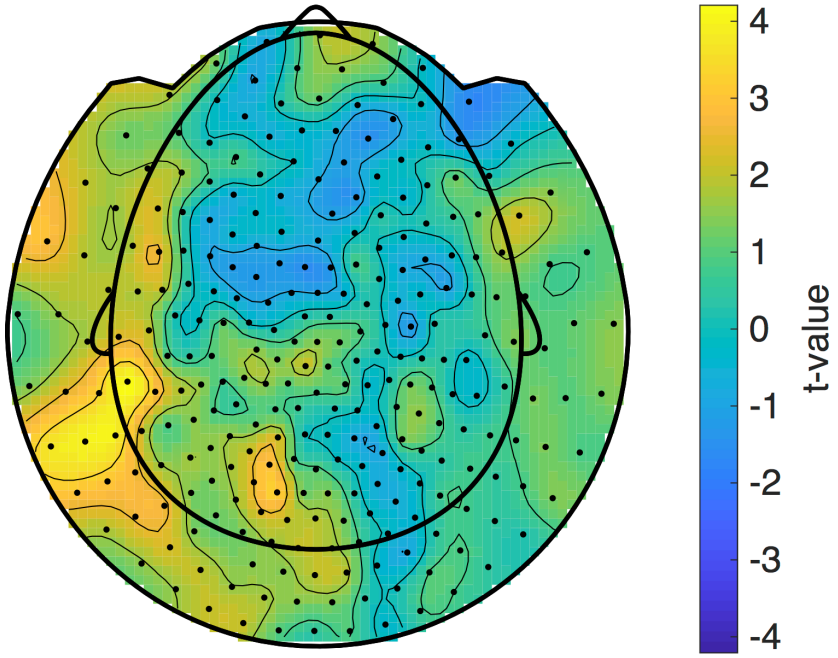
As expected, we observed a higher level of activity for scalar adjectives than for pseudoadjectives in left BA37 (p=0.017) where the cluster with the largest test statistic was between 224-281 ms. For the left BA22, significance probability was 0.034, i.e., above our significance criterion; here, the cluster with the largest test statistic was between 220-258 ms. The difference in the source-reconstructed activity in the adjective presentation window for these regions is plotted in Figure 5.8. Note that our largest temporal cluster is at a later point than in the study by Hauk and colleagues (180-220 ms). This could be due to a different task in our study, due to a different language, or due to the fact that our materials consisted of adjectives instead of nouns. Despite the difference in the temporal extent, given that we do observe an effect with the expected spatial extent and in the expected direction, this result demonstrates that our source reconstruction pipeline produces meaningful results, reflecting activity related to language processing.

For sensor-level data

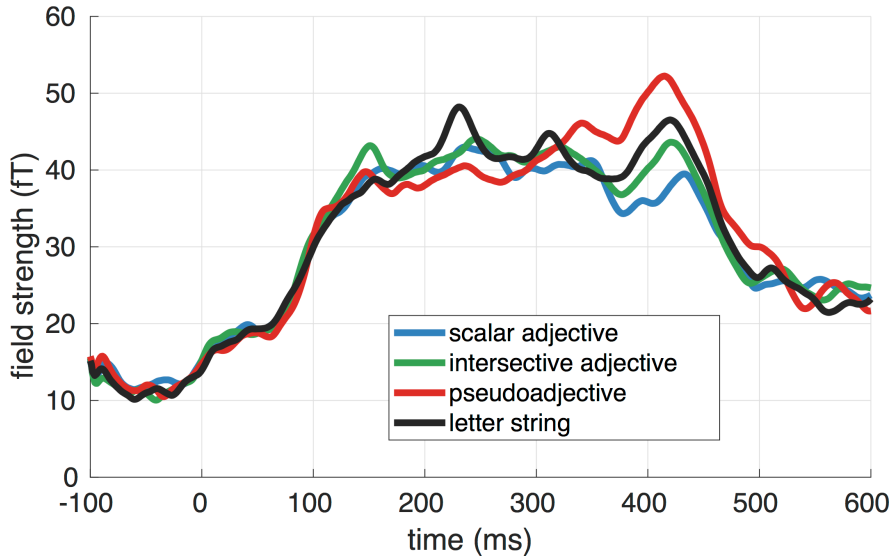
We ran a second sanity check analysis at the sensor-level, as a way to validate the sensor-level ERF analysis and, thereby, provide evidence that we can meaningfully interpret the absence of a violation effect as an indication that participants might not have performed syntactic composition in the case of pseudoadjectives. One of the robust findings in the ERP literature is an N400 (or an N400-like) effect in the lexical decision task whereby a more negative ERP is typically observed for pseudowords as opposed to real words (e.g., Barber, Otten, Kousta, & Vigliocco, 2013; Bentin, McCarthy, & Wood, 1985; Kounios & Holcomb, 1994; Meade, Grainger, & Holcomb, 2019). Consistent with this, an EEG study of minimal phrase composition which employed the same paradigm and experiment structure as the original Pylkkänen lab studies and the present study (Neufeld et al., 2016) also reported such a negativity for pseudowords, at the adjective presentation window. Given those results, we thus also expected to observe an N400-like effect for pseudoadjectives contrasted with real adjectives. We again only looked at the scalar adjectives in order to equalize word length between conditions. Note that the ERF data for this analysis was baseline-corrected using a window of 100 ms before adjective onset. We ran a paired t-test based cluster-based permutation analysis comparing ERFs, looking for spatiotemporal clusters between 100-600 ms after adjective onset. The analysis was two-sided. We used the whole time-window of adjective presentation (except for the first 100 ms) in this analysis because the N400 effects for such contrasts does not have one consistently observed time-course and is often long-lasting (Barber et al., 2013; Meade et al., 2019). The set-up of this analysis was same as for the sensor-level analysis of morphosyntactic agreement violation trials.

The cluster-based permutation test revealed a significant difference between the scalar adjectives and pseudoadjectives ($p=0.004$). The cluster with the highest test statistic was between 368-465 ms and most pronounced over left temporal sensors; see Figure 5.9a. In this time-window, signal strength was greater for pseudoadjectives than scalar adjectives (see Figure 5.9b for depiction of grand average ERFs (note that ERFs for all conditions are depicted in the plot whereas only two were analyzed). The time and spatial features of the effect are in line with what we expected based on the results of previous studies. We can thus be confident that the pre-processing pipeline of our ERF data reflects processing of the stimuli.

Figure 5.9: ERFs at the adjective presentation window. Note that the time-point 0 here corresponds to the onset of the adjective.



(a) Distribution of t-values in the time-window of the cluster with the highest test statistic: 368-465 ms after adjective onset. On this plot, the positive values correspond to greater signal strength for pseudoadjective condition as opposed to the scalar adjective condition.



(b) Grand average ERFs at the sensors above the left temporal lobe. Note that here we depict ERFs for all four adjective conditions while only pseudoadjective and scalar adjective conditions were analyzed.

5.3.6 Further exploratory analyses

In our analyses of composition effects so far, we focused on source reconstructed data following the previous MEG studies. Given that we see no differences between conditions at the source level, we investigated our data further for presence of any differences between conditions in the sensor-level data. The data at the sensor-level is noisier in comparison to source reconstructed data since multiple sources are likely contributing the signal observed at each sensor; in addition, the exact position of the participants' heads relative to the MEG sensors is not taken into account. On the other hand, doing the analyses at the sensor-level we make fewer modeling assumptions (i.e., no need to solve the inverse problem). To our knowledge, only three previous MEG studies of LATL composition effect have reported looking at the potential differences between conditions at the sensor-level, and none of them found a significant difference (Bemis & Pylkkänen, 2011; Del Prato & Pylkkänen, 2014; Zhang & Pylkkänen, 2015). Two of these studies found a difference when narrowing down the time-window for the cluster search to just 100 ms around the point where they already report a difference in the source-reconstructed neural activity (Bemis & Pylkkänen, 2011; Zhang & Pylkkänen, 2015). Nonetheless, an EEG study looking at the composition of adjective-noun phrases with the same set-up and contrast did report a difference (Neufeld et al., 2016) so it is possible that we do see some difference as well that will help us interpret our results.

We performed sensor-level analyses in a procedure parallel to the one described for the morphosyntactic control trials and sanity check analysis. For the composition effect, we looked for a difference in ERFs between a noun preceded by a real adjective as opposed to a noun preceded by a letter string between 100 and 600 ms after noun onset. Again, to equalize signal-to-noise ratio across conditions, we performed 100 iterations of this analysis with a different randomly selected subset of real adjective trials each time. Based on the cluster-based permutation analyses, there was a difference between conditions with p-value smaller than 0.05 in 85 out of 100 iterations. Because each iteration included a different subset of trials, the temporal and spatial extent of the cluster with the largest test-statistic was slightly different in each iteration. To get the most stable time-window of the cluster with the largest test statistic, we extracted the time-window that belonged to the cluster with the largest test-statistic in at least 80% of the iterations; it was 136-180 ms after noun onset. In this time-window, the difference was most pronounced on the right central, parietal and occipital sensors; on these sensors the signal strength was smaller for the real adjective condition than for the letter string condition. The distribution of summed t-values across sensors across all 100 iterations is plotted in Figure 5.10a. The ERFs for the right parietal sensors is depicted in Figure 5.10b.

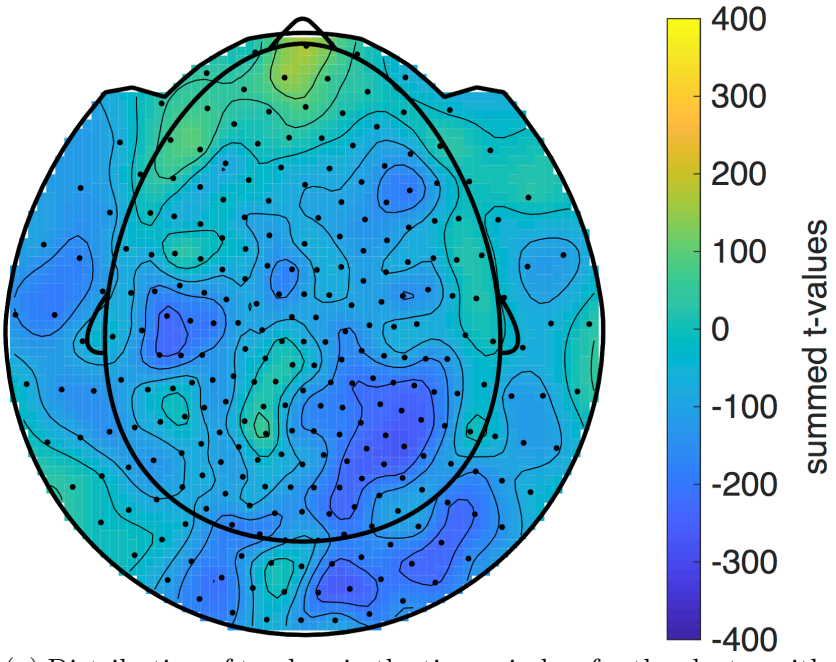
The time-window of this difference between conditions is earlier than that reported for the LATL composition effect (between approximately 200-250 ms

after noun onset for these types of materials), so it is not clear what exactly this difference in ERFs reflects. To look at potential neural sources where the difference is originating from, we plotted and inspected the source-reconstructed neural activity in the time-window of the cluster with the largest test statistic. Figure 5.10c depicts the neural activity for the whole brain in this time-window. In this time-window, there was less neural activity for real adjective trials than for letter string trials in the right and left occipital lobes. This is difficult to interpret in terms of the composition effect because we expected more neural activity for the condition which requires composition processes in comparison to the condition without composition processes. In addition, occipital areas are not considered to be strongly involved for linguistic processing, but rather responsible for low-level visual processing. Given all these reasons, we believe this difference is unlikely to be reflecting composition-related processing. Rather, it is more likely that this effect reflects differences in visual processing of the noun based on what preceded it, a real adjective or a letter string.

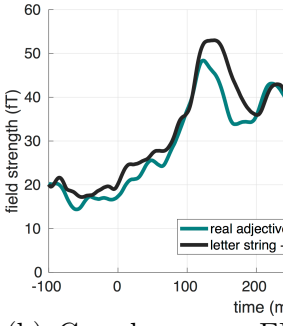
We further looked for a potential difference in ERFs between different adjective class conditions: difference between nouns preceded by an intersective adjective as opposed to nouns preceded by scalar adjectives between 100 and 600 ms after noun onset. Note that the number of trials for each condition for this analysis was equal, so there was no need to run multiple iterations with different sub-samples of trials. The cluster-based permutation analysis did not reveal any differences between conditions.

Finally, we also conducted sensor-level analyses for the syntactic composition contrast: difference between ERFs for nouns preceded by pseudoadjectives and nouns preceded by letter strings between 100 and 600 ms after noun onset. Again, the number of trials for each condition for this analysis was equal, so there was no need to run multiple iterations with different sub-samples of trials. There was a difference between conditions with a p-value of 0.02. The cluster with the largest test-statistic was in the time-window 342-424 ms, and included the central and parietal sensors in both hemispheres. In this time-window, signal strength was smaller for pseudoadjective condition compared to the letter string condition. The distribution of t-values across sensors in this time-window is depicted in Figure 5.11a. The ERFs for the right parietal sensors are depicted in Figure 5.11b. We also looked at the potential neural sources of this effect; they too seem to be located in occipital areas. The difference in source-reconstructed neural activity between these conditions in the time-window of the cluster with the largest test-statistic is depicted in Figure 5.11c. This time-window matches a time-window where a linguistic N400-like effect would appear. This effect is, however, rather weak (compare it to the effect observed in the adjective time-window in the 'sanity check' analysis). In addition, we observe less neural activity for the pseudoadjective condition where we, in fact, expected additional processing to occur, making it difficult to interpret this effect as reflecting the syntactic composition process. Finally, the localization of this effect is, again, inconsistent with

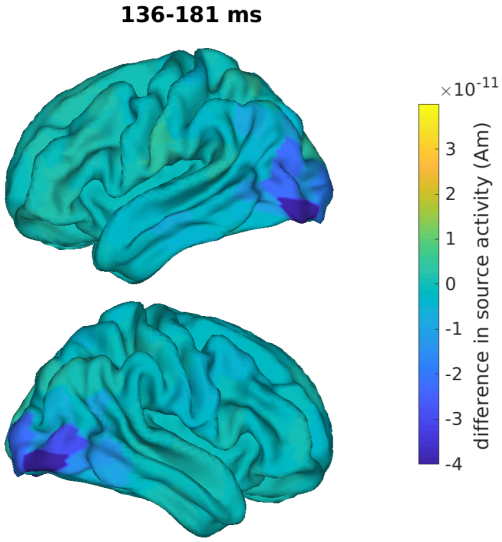
Figure 5.10: Results of the sensor-level analysis at the noun presentation window. Note that the time-point 0 here corresponds to the onset of the noun.



(a) Distribution of t-values in the time-window for the cluster with the largest test-statistic across 100 iterations: 136-180 ms after noun onset. The depicted values are summed t-statistic values across all 100 iterations (i.e., in terms of the magnitude, value 300 here corresponds to a t-statistic 3 if it was a single test). On this plot, the negative values correspond to smaller signal strength for real adjective condition as opposed to the letter string condition.



(b) Grand average ERFs on the right parietal sensors. Note that this plot depicts the ERFs from just one analysis iteration (i.e., randomly selected sub-sample of real adjective trials), and not the same one as for the neural activity plot in figure (c)

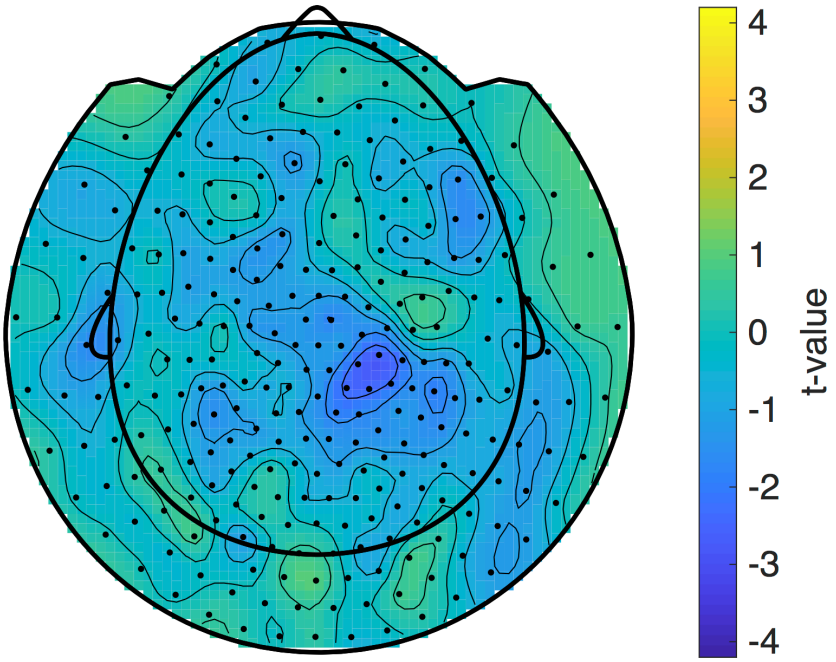


(c) Source-reconstructed activity for nouns preceded by real adjectives minus nouns preceded letter strings in the time-window selected based on the ERF analysis results. Note that this plot depicts the neural activity from just one analysis iteration (i.e., randomly selected sub-sample of real adjective trials), and not the same one as for the ERF plot in figure (b).

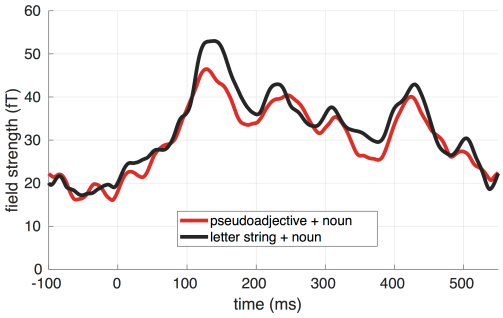
such an interpretation - occipital areas are typically considered to be responsible for low-level visual processing, and it is unclear why they would be involved here. This difference, thus, also rather reflects some other process.

Overall, we conclude that there are no clearly interpretable results related to the composition effect we are interested in from these analyses of sensor-level data. This is not surprising given that previous MEG studies looking at the composition effect at the sensor-level also failed to show such effects.

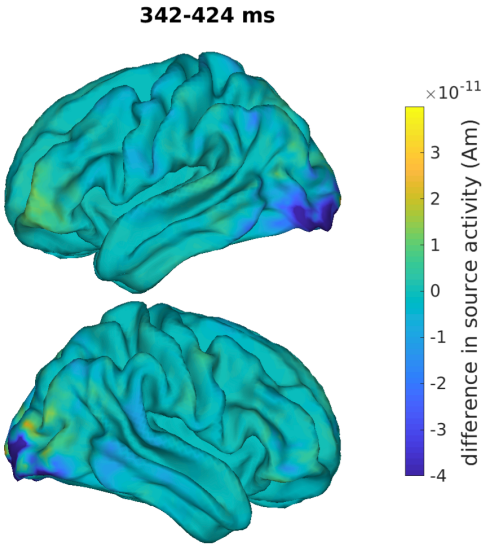
Figure 5.11: Results of the sensor-level analysis of syntactic composition: difference between nouns preceded by pseudoadjectives and by letter strings. Note that time-point 0 here corresponds to the onset of the noun.



(a) Distribution of t-values in the time-window of the cluster with the largest test-statistic: 342-424 ms after noun onset. On this plot, the negative values correspond to smaller signal strength for pseudoadjective condition as opposed to the letter string condition.



(b) Grand average ERFs on the right parietal sensors.



(c) Source-reconstructed activity for nouns preceded by pseudoadjectives minus nouns preceded by letter strings in the time-window selected based on the ERF analysis results.

5.4 General discussion

We investigated how the brain carries out minimal adjective-noun phrase composition in Dutch, using magnetoencephalography. To investigate the semantic composition processes, we followed up on previous research, which contrasted processing of an adjective-noun phrase (using adjectives of different class) with processing of a noun preceded by a meaningless string of letters (Bemis & Pylkkänen, 2011; Ziegler & Pylkkänen, 2016). We adopted a methodology parallel to the previous studies as much as possible, but conducted our study in a different language and with additional norming criteria for the materials. For part of the research questions and the corresponding analyses, we had clearly specified hypotheses based on the results of previous studies. In addition, we also conducted exploratory analyses.

To investigate syntactic composition, we introduced a novel condition to the paradigm with pseudoadjectives instead of adjectives. In this condition, we expected syntactic composition to be carried out by the participants but not semantic composition. Since there was no previous MEG study that included such a condition, we relied on the results of fMRI studies to decide about the localization of the effects that we should observe and conducted exploratory analyses.

We failed to observe a composition effect, i.e., higher levels of activity when processing nouns preceded by a real adjectives as opposed to nouns preceded by letter strings. This was the case for a confirmatory analysis with a pre-defined time-window (i.e., 200-258 ms after noun onset) and region (i.e., BA21), and for the exploratory analyses (at any point between 100-500 ms after noun onset, in any of the four selected regions based on previous studies - BA21, BA38 [temporal pole], BA39 [angular gyrus], BA44+45 [left inferior frontal gyrus] - and for any type of adjective and noun). While absence of a difference in other regions is consistent with previous studies (e.g., Bemis & Pylkkänen, 2011; Blanco-Elorrieta et al., 2018; Blanco-Elorrieta & Pylkkänen, 2016 also do not find any semantic composition effect in the left BA44+BA45 or left BA39), absence of a difference in LATL (BA21 and BA38) is striking, given that multiple previous studies reported such a difference for a parallel contrast to the one we employed in our experiment (starting from Bemis & Pylkkänen, 2011, 2013a, 2013b; though see *Table 5.5* for a detailed overview of all findings and discussion below). Accordingly, we did not observe the previously reported modulation of activity in LATL by noun specificity and adjective class. These latter results are less surprising given that only two studies to date reported a modulation by noun specificity (Westerlund & Pylkkänen, 2014; Ziegler & Pylkkänen, 2016) and only one study reported a modulation by adjective class (Ziegler & Pylkkänen, 2016) in a similar set-up. Nonetheless, absence of the basic LATL composition effect precludes us from making conclusions about these *modulating* factors. We therefore do not discuss results regarding noun specificity and adjective class further. Finally, we also fail to observe a syntactic composition effect, i.e., higher levels of activity when

processing nouns preceded by pseudoadjectives marked for grammatical gender as opposed to nouns preceded by letter strings.

Let us first discuss potential differences in data quality between ours and previous studies that could be responsible for failing to observe the previously reported LATL composition effect. Our study did not have a small number of participants or trials relative to the previous ones (our analyses included 33 participants, with 40 items per experimental design cell, whereas, for example, Z&P had 24 participants and 50 items per experimental design cell; other studies reported this effect with approximately 15-25 participants tested). Our MEG data artifact exclusion procedure is in parallel with or, perhaps, even improved in comparison to some of the previous studies (e.g., Bemis & Pylkkänen, 2011; Pylkkänen et al., 2014; Westerlund & Pylkkänen, 2014; Ziegler & Pylkkänen, 2016 etc. do not report removing heartbeat-related signal). In addition, whereas all previous studies to date (at least those we are aware of) demonstrating this effect used a scaled template brain for source activity reconstruction from MEG data, we used individual MRI scans of each participant which arguably resulted in more reliable localization of activity estimates. Finally and most importantly, in our sanity check analysis on the source reconstructed activity we replicated an independent previously reported effect; this means that the data of the present study was of adequate quality and that the pre-processing pipeline produced meaningful results. Given these considerations, we conclude that the signal-to-noise ratio in the present study was sufficient to detect the kinds of effects that have previously been shown. Therefore, it is unlikely to be the reason for the difference in the result from previous studies. Instead, we believe that content-related reasons (language, materials, task, analysis choices) should be considered.

5.4.1 Potential reasons for the failure to observe the LATL composition effect

In this section, we discuss content-related differences between our study and the previous ones reporting the LATL composition effect that could have been the reason for our null effect. Such factors may warrant further investigation to probe the limits of and relevant preconditions for observing the LATL composition effect.

One potential reason why we did not observe an LATL composition effect could be that we conducted our study in Dutch, and that the properties of composition of minimal adjective-noun phrases in Dutch are such that an LATL composition effect might not show up in an MEG study. This is unlikely, however, given that Dutch is extremely similar to English, belonging to the same language family. The only notable difference between Dutch and English adjective-noun phrases is that in Dutch the adjectives are inflected for grammatical gender: an ending ‘e’ either is or is not added to the adjectives depending on whether they

are combined with a noun belonging to one of two grammatical gender classes. This difference is unlikely to play a significant role since composition effects in left anterior temporal lobe in two-word phrases have also been reported for alternative phrase structures (specifically, for verb-noun, adverb-verb, adverb-adjective phrases in Westerlund et al., 2015, but see Kim & Pylkkänen, 2019 where only a weak effect was found and for just one class of adverbs in adverb-verb phrases) and in Arabic where adjectives also agree with nouns in terms of grammatical gender (Westerlund et al., 2015). Nonetheless, given that the evidence for the LATL composition effect is coming from only one previous study and with just one other language with grammatical gender agreement, there is a small likelihood that presence of grammatical gender agreement in Dutch could have resulted in the failure to observe the effect in the present study. If the absence of the effect in our study was indeed due to it being conducted in Dutch, it would mean that the presence of morphosyntactic agreement between an adjective and a noun somehow precludes involvement of LATL in composition. Alternatively, when morphosyntactic agreement is involved, the process is more heterogeneous than without such agreement, making it impossible to be detected. It is, for example, possible that the timing or the spatial extent of the composition process is less uniform across participants and/or nouns when morphosyntactic agreement processing is involved.

Another potential reason for the observed null effect might be a difference in stimulus materials and in the specific task we used when compared to previous studies. We reviewed all previous studies known to us that investigated the LATL composition effect, presented in *Table 5.5*. This table is not meant as an exhaustive review of all comparisons and contrasts reported in each study, but rather summarizes the results specifically for LATL ROIs with simple two-word phrases. Remember that the stimulus material of the present study consisted of scalar and intersective adjectives. When reviewing all previous studies, it is striking that, in fact, many of the studies reporting an effect used color adjectives only (7 out of 15) or color adjectives intermixed with other types of adjectives (another 3 out of 15). Only two studies reported the LATL composition effect without color adjectives in the materials (for information on the remaining three studies see footnote [22]). One of these two studies was with noun-noun phrases rather than adjective-noun phrases (Zhang & Pylkkänen, 2015), with the other one being the Ziegler and Pylkkänen (2016) study that we followed in the present study. Since only two studies observed the LATL composition without color adjectives and one study (ours) does not find evidence for such effect, it is possible that the LATL composition effect is only robustly observed for adjective-noun

[22] We are not certain that the materials in Standard Arabic used by Westerlund and Pylkkänen (2015) did not include color adjectives; the effect reported by Blanco-Elorrieta and Pylkkänen (2016) for complex numbers was much later than the expected LATL composition effect; finally, Zhang and Pylkkänen (2018) report significant effects for other types of adjectives, but only when analyzed separately and without multiple comparisons correction reported.

phrases with color adjectives. If this is the case, it would mean that LATL is particularly sensitive to composition with color adjectives or that the composition process is more uniform across participants and/or nouns with color adjectives.

A similar line of reasoning is applicable to the specific task that participants had. In the present study, participants read a phrase and were asked to respond to a comprehension question at the end of each trial, with the question being about the combined adjective and noun meaning. As can also be seen in *Table 5.5*, the LATL composition effect has been reported predominantly in studies where participants matched a phrase to a picture or produced a phrase as a description of a picture. In contrast, only a small portion of the previous studies that observed the LATL composition effect reliably (4 out of 15) used a simple text-based comprehension task comparable to the task used in our study. Possibly, the presence of a picture in the task leads to participants being more likely to engage in imagery than a text-based comprehension task. Thus, we believe our null result then tentatively suggests that perhaps the LATL is more robustly involved in composition when participants have to engage in imagery of the objects[23]. The potential dependence of the LATL composition effect on the specific task used should therefore be investigated in future research.

A striking aspect of previous studies concerns the variability in the regions of interest that were used in the analyses. As already noted in the *Introduction* section, while previous studies always referred to the region in which the composition effect was observed as the "left anterior temporal lobe", its exact spatial extent did differ, sometimes dramatically, ranging from the anterior to the posterior portions of the left temporal lobe. Different studies defined the LATL as (all left) BA21, as BA38, as BA38+BA20+BA21 combined, as BA38+BA20+BA21 separately, as BA38 and the anterior portions of BA20 and BA21, by using spatial clustering in and around LATL or in the whole temporal lobe. The spatial extent of the observed effect is likely to differ somewhat in different studies given the limited spatial resolution provided by source reconstruction based on MEG data. However, given the breadth of definitions of the region of interest where an effect was reported, it is difficult to conclude with confidence that these are not, in fact, many different effects that were grouped under the same umbrella, or that at least some of these effects are not false positives[24]. To avoid these issues, in the present study we preferred to only conduct confirmatory analyses in exactly the same region of interest as the original study with the same design that reported the effect, with other regions being examined only in an exploratory way.

From a more general perspective, the present and previous studies assume that an identical cortical region (i.e., set of dipoles) across participants will perform an

[23]This, speculatively, is in line with the potential dependence on color adjectives being used since color adjectives are potentially more likely to involve imagery than other types of adjectives (compare: "white guitar" vs "large guitar").

[24]Note that this point of concern is also applicable to the differing time-windows that were analyzed for the presence of the effect across different studies.

identical functional operation across participants – composition of an adjective and a noun. This assumption may not be warranted given that different (but probably neighbouring) cortical regions might be responsible for composition in different participants. This would in turn lower the strength of the signal in the averaged data (it should be noted that given the already low limited spatial resolution of MEG source reconstruction, small differences may not play a substantial role, but there would nonetheless be some reduction in the strength of the observed signal, depending on the extent of variability). An alternative and perhaps a better long-term solution for this line of research (solving the problem of researcher degrees of freedom with ROI definition described above and doing away with the assumption of identical cortical regions performing composition) would be to use a functional localizer for ROI definition. Such a functional localizer would use a pre-defined set of materials that would be identical across studies and potentially localize the ROI on individual subject level (note that Flick et al., 2018 already used a functional localizer for ROI definition, but did so on a group level and with a novel set of materials; our suggestion to use an individual-level functional localizer is in fact rather ambitious given that many trials would need to be administered to each participant to reach a reasonable signal-to-noise ratio). Such a functional localization on an individual level before running group-level analyses has been previously proposed for fMRI studies (Fedorenko, Hsieh, Nieto-Castañón, Whitfield-Gabrieli, & Kanwisher, 2010). In other words, the region of interest can be defined as the region showing increased activity for nouns preceded by real adjectives as compared to nouns preceded by letter strings for a particular participant given the same set of materials across studies. Whether different factors such as noun specificity and adjective class modulate the composition effect can then be investigated within this functionally defined ROI.

Overall, we conclude that our failure to observe the widely reported LATL composition effect is not as surprising after carefully considering the specifics of materials, tasks and regions of interest used in previous studies where such an effect *was* reported. Our review and discussion suggest that further investigation of limits and relevant preconditions for observing the LATL composition effect is needed.

Table 5.5: Summary of findings of studies that investigated composition of minimal two-word phrases using MEG. This table summarizes the results for LATL ROI analyses only. Only those comparisons are included in the table that are relevant to the presence/absence of the simple LATL composition effect (e.g., Flick et al., 2018 looked at compound nouns combined with adjectives which we omitted here).

Study	**Task**	**Composition condition materials** (language is English if not specified otherwise)	**Control/no composition condition**	**LATL ROI definition**	**Time-window where an effect was looked for** (after noun onset; using cluster-based permutation test if not specified otherwise)	**Effect observed at $p<0.05$?**	**Time-window of the cluster with the largest test-statistic** (after noun onset if not specified otherwise)
Bemis & Pylkkänen, 2011	picture-matching task	color adjective + noun	consonant string + noun; additional control: making sure that this difference is larger than the difference between a two-word list condition ["boat,cup"] and a one word condition ["cup"]	the area around BA 38 and the anterior portions of BA 20 and 21	0-500 ms	yes	184-255 ms
Bemis & Pylkkänen, 2013a	picture-matching task	color adjective + noun	consonant string + noun; additional control: making sure that this difference is larger than the difference between a two-word list condition ["boat,cup"] and a one word condition ["cup"]	based on a group-level full-brain analysis looking for sources that showed composition activity in both visual and auditory modalities at any point in time between 0-600 ms; this analysis identified a 'cluster within the LATL'	unspecified (likely 0-600 ms)	yes	191-299 ms (visual phrase presentation); 268-323 ms (auditory phrase presentation, p=0.086)
Bemis & Pylkkänen, 2013b	picture-matching task	color adjective + noun; noun + color adjective (reversed order)	consonant string + noun; consonant string + adjective	the area around BA 38 and the anterior portions of BA 20 and 21	0-600 ms	yes	201-269 ms
Westerlund & Pylkkänen, 2014	picture-matching task	mix of color and other adjectives + low specificity noun	consonant string + noun	BA38 , BA20, BA21 combined	200-400 ms	yes	218-282 ms
		mix of color and other adjectives + high specificity noun				no	-
Pylkkänen et al., 2014	production: picture-naming task	color adjective + noun	listing two colors ("red", "blue")	the area around BA 38 and the anterior portions of BA 20 and 21	unspecified (likely 0-700 ms)	yes	248-410 ms after picture onset
			naming an object ("star")		0-500 ms	yes	175-475 ms after picture onset
			naming a color ("red")		0-500 ms	yes	266-500 ms after picture onset
Del Prato & Pylkkänen, 2014	production: picture-naming task	color adjective + noun	listing two colors ("red", "blue")	BA38	100-400 ms	yes	212-400 ms after picture onset

Westerlund et al., 2015	comprehension question	number word + noun ("two cups")	mix of color and other adjectives + noun; adverb + verb; adverb+adjective; verb+noun; preposition+noun; determiner+noun	listing two numbers ("two, three")	consonant string + noun/verb/adjective	BA38, BA20, BA21 combined	0-600 ms	yes	-
			in Modern Standard Arabic: adjective (unclear if color adjectives were included) + noun; noun + adjective; verb + noun	consonant string + adjective/noun				yes	35-453 ms
Zhang & Pylkkänen, 2015	comprehension question	noun + noun ("tomato soup")		consonant string + noun	BA38		150-350 ms	yes	156-348 ms
Blanco-Elorrieta & Pylkkänen, 2016	production: picture-naming task	color adjective + number word ("green threes")		listing two numbers ("two, three")	BA38, BA20, BA21 separately, FDR-correction for MCP at rate 0.1 reported		150-400 ms	yes	216-246 ms
								yes	150-258 ms after picture onset in BA20;
									184-349 ms after picture onset in BA38
			complex number ("twenty-three")					yes	435-589 ms after picture onset in BA21
			number word + number ("two threes")					no	-
Ziegler & Pylkkänen, 2016	comprehension question	scalar adjective + noun		consonant string + noun	BA21		200-500 ms	no	-
		intersective adjective + noun						yes (but only for intersective adjectives combined with low specificity nouns)	200-258 ms

Neufeld et al., 2016 (EEG study)	picture-matching task	color adjective + noun	consonant string/pseudoword + noun	mean value of EEG electrodes grouped by anteriority: anterior, posterior	mean value between 184-256 ms	yes	184-256 ms
Blanco-Elorrieta et al., 2018	production: picture-naming task	in American Sign Language: color adjective + noun	background color + noun ("blue, lamp")	described as "LATL (including BA38)"	100-300 ms	yes	145-189 ms after picture onset
		color adjective+ noun				yes	221-300 ms after picture onset
Zhang & Pylkkänen, 2018	picture-matching task	color adjective / "same" / "different" / "larger" / "smaller" + noun	"another" + noun	group-level spatiotemporal clustering for each modifier type within the area comprised of: temporal pole, entorhinal, parahippocampal, fusiform, and superior, middle, and inferior temporal gyri.	180-250 ms	no	185-240 ms reported but not significant; pairwise comparisons for each type of modifier are reported with significant differences from the control condition but no MCP correction reported.
Flick et al., 2018	comprehension question	mix of color and other adjectives + noun	consonant string + noun	group-level spatiotemporal clustering within the left temporal lobe (subsequently the obtained spatial cluster was used as a functional ROI for other analyses)	150-350 ms	yes	162-299 ms
Kim & Pylkkänen, 2019	comprehension question	(eventive, resultative or agentive) adverb + verb	consonant string + verb	BA20, BA21, BA38 separately (FDR-correction within hemisphere for MCP reported)	200-300 ms	no	280-296 ms in BA38 (for eventive adverbs only) reported but did not survive MCP correction

5.4.2 Syntactic composition effect

In order to investigate syntactic composition of adjective-noun phrases, we included a condition in which a pseudoadjective was combined with a noun. We expected that participants would carry out syntactic composition in these phrases without carrying out semantic composition since the pseudoadjective had an inflection that agreed with the grammatical gender of the noun. However, in principle our participants could also simply have ignored the pseudoadjectives and still respond correctly to the comprehension question (since it was only about the noun in this condition). As a way to check whether our participants did or did not pay attention to the pseudoadjectives, we added trials where the grammatical gender marking of the pseudoadjective and the grammatical gender of the noun did not match (i.e., syntactic violations). We expected to observe a grammatical violation signature in ERFs during processing of the noun in the form of an N400-like effect. Such an effect has been observed for grammatical gender agreement violations in adjective-noun phrases in other languages (Molinaro et al., 2011) and for grammatical gender agreement violations in determiner-noun phrases in Dutch (Hagoort, 2003). However, we did not observe such an effect, meaning that either participants did not perform syntactic composition in this condition or that we did not have enough power in this part of the study to observe a violation detection effect. The latter could indeed have been the case as we used only 20 trials with matching and 20 trials with mismatching pseudodjective inflection for this control analysis; we deliberately did so as we wanted to avoid fatiguing our participants with a longer experiment (note that this part of the study was administered at the very end of the experimental session). It is thus possible that the signal-to-noise ratio was not sufficiently high in this analysis for an effect to emerge. In a follow-up study, we would choose to administer more trials for this control analysis in order to be confident that we are not missing the violation detection signature. Another complication of our control trials concerns the fact that we could analyze data only up to 600 ms after noun onset (300 ms of the noun presentation followed by 300 ms of a blank screen after which a comprehension question was displayed). This trial structure was chosen in order to have it identical to the trial structure in the main part of the experiment on the LATL composition effect (and, therefore, to make sure that participants do not notice a difference and switch to a different mode of processing). But this implies that we cannot analyse the data of this control condition beyond 600 ms after noun onset and thus we cannot look at the time-window of a P600-like effect (approximately 500-800 ms after noun onset) which has sometimes been observed in addition or instead of an N400-like effect for similar grammatical violations (Hagoort & Brown, 1999; Molinaro et al., 2011). In a future study, it might be reasonable to have a longer blank screen after noun presentation in order to be able to look at the P600-like time-window as well.

Turning to the syntactic composition itself, we do not observe any syntactic

composition effect, i.e., difference between processing nouns preceded by pseudoadjectives and nouns preceded by letter strings. It is possible that we do not observe such a difference because our participants simply did not perform syntactic composition (given the results with our control trials). It is, however, also possible that they *did* perform syntactic composition, but we were not able to detect it with our method. As discussed in the *Introduction*, syntactic composition effects for minimal phrases have previously been reported in fMRI studies (Schell et al., 2017; Zaccarella & Friederici, 2015; Zaccarella et al., 2017), but there were no MEG studies specifically investigating it. It is possible that the time-course of syntactic composition is too variable between participants or phrases making it difficult to detect with averaged data with high temporal resolution as in the case of MEG, but possible with BOLD data which is capturing activity averaged over a longer time-span. Alternatively, spatial resolution of MEG source-reconstructed data is not good enough to capture the syntactic composition-related activity (for example, Zaccarella & Friederici, 2015 identify just a part of BA44 as most correlated with syntactic composition, a rather small patch of cortex). Even with fMRI data, however, syntactic composition has not always been observed for minimal phrases (Matchin et al., 2017). In this context, some have suggested that syntactic composition of a simple minimal phrase like the one used here is highly automatic, hence not needing many resources to be processed and not detectable (e.g., by Flick & Pylkkänen, 2018; Matchin et al., 2017; Pylkkänen, 2019).

In summary, it remains unclear based on our findings whether the brain is insensitive to 'pure' syntactic composition or whether the methods we employed here are simply not appropriate for detecting such an effect. Moreover, given the absence of a basic composition effect there are no straightforward conclusions to be drawn about potential differences between composition involving semantics on the one hand, and 'purely' syntactic composition on the other.

5.5 Conclusion

The present study investigated composition of minimal adjective-noun phrases in Dutch with two goals. The first goal was to look for a previously established composition effect in left anterior temporal lobe, as well as its modulation by noun specificity and adjective class; the latter modulation has been taken as evidence for the semantic nature of this composition effect. The second goal was to target specifically syntactic composition processes by including a novel condition where a pseudoadjective's inflection matched the noun in terms of grammatical gender.

We failed to observe the previously reported LATL composition effect. Our review of previous studies that reported the LATL composition effect suggests that most likely this effect is only robustly observed when materials consist of color adjectives and/or include an imagery task. Thus, in our view, future research should focus on limits and relevant preconditions for observing the LATL

composition effect. Additionally, our review reveals substantial inconsistencies in previous studies in terms of the brain regions and time-windows of interest that were used for analyses. We argue for more consistent definitions of regions and time-windows of interest in this line of research.

We did not observe a specifically syntactic composition effect either. However, because our control condition did not show with certainty that participants engaged in syntactic composition of pseudoadjective-noun phrases, we cannot be sure that the failure to observe this effect is not simply due to participants' failure to engage in syntactic composition in this condition. We acknowledge that our control condition should have been statistically better powered, which we hope follow-up studies in a similar vein will ensure.

To conclude, the study of semantic (and syntactic) composition processes in minimal two-word phrases is surely a promising area of research as it allows one to study such composition processes in tightly controlled experimental settings. However, the present data also suggest that these tightly controlled settings should be used systematically to trace potential influences of the specific linguistic materials and tasks on composition effects before any general conclusions about semantic (and syntactic) composition can be drawn.

5.6 Data Accessibility

Stimuli, preprocessed data, and analysis scripts are available on https://osf.io/kyc4u/, DOI 10.17605/OSF.IO/KYC4U. Raw data in MEG-BIDS format is currently under preparation for public release.

Chapter 6

Summary and avenues for future research

Each of the chapters in this thesis contains a general discussion section where the findings and their implications are discussed in detail. In this concluding chapter, I give a short summary of each study and highlight some of the most promising avenues for future research which were opened or emphasized by these studies.

6.1 Quantifiers and number symbols as symbolic references to magnitude information

Both natural language quantifiers and number symbols are symbolic references to quantity information, so we reasoned that it would be useful to consider their relationship to cognitive mechanisms for nonsymbolic quantity processing. In Chapter 2, we first reviewed existing research on the developmental, behavioral, and neuronal aspects of number symbol processing in relation to nonsymbolic quantity processing mechanisms with which humans are equipped. We used findings and paradigms from that line of research to discuss research on the potential parallelism between natural language quantifiers and nonsymbolic quantity processing mechanisms. This approach allowed us to formulate a number of new research questions regarding quantifier processing and to suggest paradigms that could be used to investigate them.

From the developmental perspective, there are some promising findings regarding the potential relationship between quantifier acquisition and maturing nonsymbolic quantity processing mechanisms that should be followed up in the future. We suggest that these questions should be investigated using correlational and training studies. The behavioral paradigm that has been used in a relatively large number of past studies on the interface between quantifiers and nonsymbolic quantity processing, both in children and adults, is sentence-picture verification. To date, only proportional quantifiers have been investigated in this paradigm, whereas we suggest that such an interface could exist for other quantifier classes

as well. Comparison of what kind of information different quantifier classes extract from visual scenes would provide a useful insight into the differences between the processing mechanisms involved for each type of quantifier. We highlighted that neuroimaging studies have only investigated a small set of quantifiers so far, and with small sample sizes, and thus we can not draw strong conclusions from these studies. Importantly, future research should focus on comparing the neuronal populations involved in the processing of quantifiers and lexical items that do not refer to quantities. Ideally, these future studies should also identify the neuronal populations involved in quantity processing from perceptual input within the same data collection in order to be able to make direct comparisons with those involved in quantifier processing.

6.2 Scalar adjectives as symbolic references to magnitude information

Quantifiers can be seen as symbolic references to the quantity information represented in our cognitive systems for processing numerical magnitude. Similarly scalar adjectives can be seen as symbolic references to magnitude information in other dimensions. Whereas Chapter 2 discussed the potential connection between processing natural language quantifiers and nonsymbolic quantity processing, in Chapter 4 we formulated and tested a related hypothesis for scalar adjectives. Specifically, we hypothesized that the generalized analog magnitude representation system is recruited when the meaning of scalar adjectives is retrieved and processed. The generalized analog magnitude representation system (GMS) has been proposed as a magnitude representation and processing mechanism shared between different dimensions. Here we again took the existing research into number symbol processing as starting point. Previous studies have demonstrated the involvement of GMS representations when the meaning of number symbols is retrieved. We ran follow-up experiments to those on number symbol processing and novel experiments with scalar adjectives using the same paradigms. In the critical experiments, participants made judgments about the meaning or physical sizes of antonymous pairs of adjectives. Although our series of experiments did not provide support for our "GMS-hypothesis" regarding scalar adjectives, the hypothesis and its variations remain interesting and can be investigated in a number of other ways in future.

In our experiments, we looked at whether GMS is recruited in the retrieval of the meaning of scalar adjectives presented in isolation, i.e. out of context. In follow-up studies, this specific question could be investigated using a subliminal priming paradigm that previously provided strong evidence for recruitment of GMS in the case of Arabic digits (Lourenco et al., 2016). Furthermore, it is possible that scalar adjectives only recruit GMS representations when they are used in a phrasal context. After all, scalar adjectives can only be interpreted in

a meaningful way when they are combined with specific nouns. To test this prediction, future studies should investigate GMS recruitment in a paradigm where the meaning of the adjective has to be retrieved within a phrase or a sentence.

The hypothesis we put forward also predicts that processing scalar adjectives (but not non-gradable adjectives) should involve neuronal populations of which we know that they are involved in the processing of magnitudes from perceptual input. In fact, this prediction can in principle be tested with the MEG data presented in Chapter 5.[1] There, participants read and interpreted adjective-noun phrases with scalar and non-gradable adjectives. Thus, we would expect the involvement of these neuronal populations at some point during reading of noun-phrases with of scalar adjectives, but not during reading noun-phrases with non-gradable adjectives.

The approach suggested here can also be used to look into the processing of scalar adjectives referring to a non-perceptual dimension. While here we focused on scalar adjectives describing a perceptual dimension, there are also scalar adjectives referring to more abstract properties — e.g., 'easy', 'difficult', 'kind', 'cruel', 'happy', etc. Furthermore, our discussion has focused on scalar adjectives referring to one dimension only, but there are also scalar adjectives that can be argued to refer to magnitudes along multiple dimensions simultaneously, such as 'healthy', 'intelligent', 'typical', etc. (Sassoon, 2013). These other scalar adjectives clearly play an important role in everyday language as well, and, intuitively, the neurocognitive processes behind them should overlap with those for scalar adjectives referring to perceptual dimensions. We considered one-dimensional adjectives referring to perceptual dimensions as a more basic form, a starting point which can be used to look at the processing of these other adjectives as well.

6.3 The context-sensitivity of scalar adjectives reflected in language composition at the neuronal level

Chapter 5 focused on the context-dependence of scalar adjectives. Specifically, we focused on the dependence of the meaning of the scalar adjective on the noun with which it is combined. Participants read scalar adjectives or non-gradable adjectives combined with nouns and answered a question about the meaning of the composed phrase. Previously, a number of similar studies reported a basic composition effect in the level of neural activity in the left anterior temporal lobe (LATL). This basic composition effect refers to a difference in the neural activity between an adjective-noun phrase and a control condition (a "phrase" consisting

[1]The raw data collected in this study is publicly available, so other researchers can conduct additional analyses.

of a letter string and a noun). Ziegler and Pylkkänen (2016) in addition reported a modulation of this basic composition effect by the type of adjective (scalar adjectives versus non-gradable adjectives). This latter difference in the LATL activity was attributed to the fact that a threshold (given the comparison class) needs to be computed for a scalar adjective, but not for a non-gradable adjective. The goal of our study was, in a first step, to establish whether this difference is robust and observable in a similar set-up in a different lab and language. However, we did not observe the basic composition effect in our study and, subsequently, no modulation of this effect by the adjective class either. Reviewing in detail the previous studies that reported the basic composition effect, we put forward a number of hypotheses regarding potential factors that may be important for the observation of the basic composition effect in LATL. These factors have not been taken into consideration in past. Furthermore, we noted that the statistical analyses conducted in these past studies could have possibly led to false positive findings in some cases.

The basic composition effect in LATL was a prerequisite to observe any difference between scalar and non-gradable adjectives. Hence, we believe future research should first identify the conditions under which the basic composition effect in LATL is robustly observable (see also the discussion section of chapter 5). Once these conditions are identified, one can again address the question whether there are any differences in neural processing signature between scalar and non-gradable adjectives.

While Ziegler and Pylkkänen present the plausible and appealing suggestion that the difference in the time-course of neural activity in LATL when processing scalar adjective-noun phrases versus non-gradable adjective-noun phrases reflects a difference in the requirement of computation of a threshold, this observation alone is not sufficient to accept this theoretical interpretation in terms of a specific underlying cognitive process. If existence of such difference in neural activity is confirmed in future studies, further investigation will be needed to offer convincing support that this difference arises due to the requirement to compute a threshold. This is what we originally found appealing about the findings of Ziegler and Pylkkänen — if this difference is real and robust, one could follow it up with more detailed investigations of whether it was the threshold computation that was driving the LATL activity. Specifically, follow-up experiments could manipulate the experimental task or the context in which the phrase occurs in such a way that in one case computation of a threshold would be necessary for a particular adjective-noun phrase and in the other case such computation would not be necessary for the same adjective-noun phrase (or threshold computation could be made more or less difficult). We would then expect to observe a modulation of LATL activity by this manipulation.

6.4 Closing remarks

This thesis looked into the cognitive and neuronal mechanisms of processing scalar adjectives and quantifiers from two perspectives: from the perspective of processing mechanisms for perceptually assessed magnitude and from the perspective of the context-dependence of their meaning. It has presented extensive discussions and suggestions for future research on the relationship between scalar adjectives and quantifiers on the one hand and perceptual magnitude processing on the other hand. Because this is a novel perspective, clearly development of a more fleshed-out theory of this potential connection is needed to make further progress. Nonetheless, I hope the studies presented here can function as a starting point for this line of research and for future developments. The study looking at the neural correlates of semantic composition has contributed in a different way — it tested the robustness of a paradigm that promises to be a great avenue for future research. Resolving the issues that this study highlighted will undoubtedly move research on the neural processes of semantic composition forward and offer an opportunity for connecting insights from research in formal semantics with research in cognitive neuroscience of language.

References

Algom, D., Dekel, A., & Pansky, A. (1996). The perception of number from the separability of the stimulus: The stroop effect revisited. *Memory & Cognition*, *24*(5), 557–572. doi: 10.3758/BF03201083

Alxatib, S., & Sauerland, U. (2019). Vagueness. In C. Cummins & N. Katsos (Eds.), *The Oxford Handbook of Experimental Semantics and Pragmatics.* Oxford University Press. doi: 10.1093/oxfordhb/9780198791768.001.0001

Anobile, G., Cicchini, G. M., & Burr, D. C. (2016). Number as a primary perceptual attribute: A review. *Perception*, *45*(1-2), 5–31. doi: 10.1177/0301006615602599

Ansari, D. (2007). Does the Parietal Cortex Distinguish between "10," "Ten," and Ten Dots? *Neuron*, *53*(2), 165–167. doi: 10.1016/j.neuron.2007.01.001

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2019). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*. doi: 10.3758/s13428-019-01237-x

Aparicio, H., Xiang, M., & Kennedy, C. (2016). Processing gradable adjectives in context: A visual world study. In *Semantics and Linguistic Theory* (Vol. 25, pp. 413–432).

Arend, I., & Henik, A. (2015). Choosing the larger versus choosing the smaller: Asymmetries in the size congruity effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1821–1830. doi: 10.1037/xlm0000135

Arsalidou, M., & Taylor, M. J. (2011). Is 2+2=4? Meta-analyses of brain areas needed for numbers and calculations. *NeuroImage*, *54*(3), 2382–2393. doi: 10.1016/j.neuroimage.2010.10.009

Ash, S., Ternes, K., Bisbing, T., Min, N. E., Moran, E., York, C., ... Grossman, M. (2016). Dissociation of quantifiers and object nouns in speech in focal neurodegenerative disease. *Neuropsychologia*, *89*, 141–152. doi: 10.1016/j.neuropsychologia.2016.06.013

Assadollahi, R., & Pulvermüller, F. (2003). Early influences of word length and

frequency: A group study using MEG. *NeuroReport*, *14*(8), 1183. doi: 10.1097/00001756-200306110-00016

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2019). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *arXiv:1902.06122 [q-bio, stat]*.

Bar, H., Fischer, M. H., & Algom, D. (2019). On the linear representation of numbers: Evidence from a new two-numbers-to-two positions task. *Psychological Research*, *83*(1), 48–63. doi: 10.1007/s00426-018-1063-y

Barber, H. A., Otten, L. J., Kousta, S.-T., & Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and Language*, *125*(1), 47–53. doi: 10.1016/j.bandl.2013.01.005

Barner, D. (2012). Bootstrapping Numeral Meanings and the Origin of Exactness. *Language Learning and Development*, *8*(2), 177–185. doi: 10.1080/15475441.2012.635541

Barner, D., Chow, K., & Yang, S.-J. (2009). Finding one's meaning: A test of the relation between quantifiers and integers in language development. *Cognitive Psychology*, *58*(2), 195–219. doi: 10.1016/j.cogpsych.2008.07.001

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: 10.1016/j.jecp.2008.12.001

Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2015). QRTEngine: An easy solution for running online reaction time experiments using Qualtrics. *Behavior Research Methods*, *47*(4), 918–929. doi: 10.3758/s13428-014-0530-7

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi: 10.1016/j.jml.2012.11.001

Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition*, *86*(3), 201–221. doi: 10.1016/S0010-0277(02)00178-6

Barth, H., La Mont, K., Lipton, J., & Spelke, E. S. (2005). Abstract number and arithmetic in preschool children. *Proceedings of the National Academy of Sciences*, *102*(39), 14116–14121. doi: 10.1073/pnas.0505512102

Barwise, J., & Cooper, R. (1981). Generalized Quantifiers and Natural Language. *Linguistics and Philosophy*, *4*(2), 159–219.

Bastiaansen, M. C., & Knösche, T. R. (2000). Tangential derivative mapping of axial MEG applied to event-related desynchronization research. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *111*(7), 1300–1305. doi: 10.1016/s1388-2457(00)00272-8

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). doi:

10.18637/jss.v067.i01

Bates, D., & Maechler, M. (2019). *Matrix: Sparse and dense matrix classes and methods*.

Bemis, D. K., & Pylkkänen, L. (2011). Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *The Journal of Neuroscience*, *31*(8), 2801–14. doi: 10.1523/JNEUROSCI.5003-10.2011

Bemis, D. K., & Pylkkänen, L. (2013a). Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cerebral Cortex*, *23*(8), 1859–1873. doi: 10.1093/cercor/bhs170

Bemis, D. K., & Pylkkänen, L. (2013b). Flexible composition: MEG evidence for the deployment of basic combinatorial linguistic mechanisms in response to task demands. *PLoS ONE*, *8*(9), e73949. doi: 10.1371/journal.pone.0073949

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology*, *60*(4), 343–355. doi: 10.1016/0013-4694(85)90008-2

Besner, D., & Coltheart, M. (1979). Ideographic and alphabetic processing in skilled reading of English. *Neuropsychologia*, *17*(5), 467–472. doi: 10.1016/0028-3932(79)90053-8

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–536. doi: 10.1016/j.tics.2011.10.001

Birnbaum, M. H. (2000). Introduction to psychological experiments on the internet. In M. H. Birnbaum (Ed.), *Psychological Experiments on the Internet* (p. xv-xx). San Diego: Academic Press. doi: 10.1016/B978-012099980-4/50001-0

Blanco-Elorrieta, E., Kastner, I., Emmorey, K., & Pylkkänen, L. (2018). Shared neural correlates for building phrases in signed and spoken language. *Scientific Reports*, *8*(1), 5492. doi: 10.1038/s41598-018-23915-0

Blanco-Elorrieta, E., & Pylkkänen, L. (2016). Composition of complex numbers: Delineating the computational role of the left anterior temporal lobe. *NeuroImage*, *124*, 194–203. doi: 10.1016/j.neuroimage.2015.08.049

Bonn, C. D. (2015). *On Theories of Abstract, Quantitative Representation* (PhD Thesis). University of Rochester, New York.

Bonn, C. D., & Cantlon, J. F. (2012). The origins and structure of quantitative concepts. *Cognitive neuropsychology*, *29*(1-2), 149–73. doi: 10.1080/02643294.2012.707122

Bonn, C. D., & Cantlon, J. F. (2017). Spontaneous, modality-general abstraction of a ratio scale. *Cognition*, *169*, 36–45. doi: 10.1016/J.COGNITION.2017.07.012

Bornkessel, I., Zysset, S., Friederici, A. D., von Cramon, D. Y., & Schlesewsky, M. (2005). Who did what to whom? The neural basis of argument hierarchies

during language comprehension. *NeuroImage*, *26*(1), 221–233. doi: 10.1016/j.neuroimage.2005.01.032

Bowern, C., & Zentz, J. (2012). Diversity in the Numeral Systems of Australian Languages. *Anthropological Linguistics*, *54*(2), 133–160. doi: 10.1353/anl.2012.0008

Brannon, E. M., & Terrace, H. S. (1998). Ordering of the Numerosities 1 to 9 by Monkeys. *Science*, *282*(5389), 746–749. doi: 10.1126/science.282.5389.746

Brants, T., & Franz, A. (2009). *Web 1T 5-gram, 10 European Languages Version 1 LDC2009T25. Web Download.* Philadelphia: Linguistic Data Consortium.

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, *120*(2), 163–173. doi: 10.1016/j.bandl.2010.04.002

Brennan, J., & Pylkkänen, L. (2017). MEG Evidence for Incremental Sentence Composition in the Anterior Temporal Lobe. *Cognitive Science*, *41 Suppl 6*, 1515–1531. doi: 10.1111/cogs.12445

Breukelaar, J. W. C., & Dalrymple-Alford, J. C. (1998). Timing ability and numerical competence in rats. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*(1), 84–97. doi: 10.1037/0097-7403.24.1.84

Brief Guide to Stan's Warnings. (2020). https://mc-stan.org/misc/warnings.html#tail-ess.

Brysbaert, M. (2019). *Power analysis and effect size in mixed effects models: A tutorial.* http://crr.ugent.be/archives/2014.

Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, *1*(1), 9. doi: 10.5334/joc.10

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3), 441–458. doi: 10.1037/xhp0000159

Buckley, P. B., & Gillman, C. B. (1974). Comparisons of digits and dot patterns. *Journal of Experimental Psychology*, *103*(6), 1131–1136. doi: 10.1037/h0037361

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, *6*(1), 3–5. doi: 10.1177/1745691610393980

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An Evaluation of Amazon's Mechanical Turk, Its Rapid Rise, and Its Effective Use. *Perspectives on Psychological Science*, *13*(2), 149–154. doi: 10.1177/1745691617706516

Bulthé, J., De Smedt, B., & Op de Beeck, H. (2014). Format-dependent representations of symbolic and non-symbolic numbers in the human cortex as revealed by multi-voxel pattern analyses. *NeuroImage*, *87*, 311–322. doi: 10.1016/J.NEUROIMAGE.2013.10.049

Bulthé, J., Smedt, B. D., & de Beeck, H. P. O. (2015). Visual Number Beats Abstract Numerical Magnitude: Format-dependent Representation of Arabic Digits and Dot Patterns in Human Parietal Cortex. *Journal of Cognitive Neuroscience*, *27*(7), 1376–1387. doi: 10.1162/jocn_a_00787

Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411.

Callaway, E. (2013). Dyscalculia: Number games. *Nature News*, *493*(7431), 150. doi: 10.1038/493150a

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., ... Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. doi: 10.1038/s41562-018-0399-z

Campbell, J. I. D., & Epp, L. J. (2004). An Encoding-Complex Approach to Numerical Cognition in Chinese-English Bilinguals. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *58*(4), 229–244. doi: 10.1037/h0087447

Cantlon, J. F., & Brannon, E. M. (2006). Shared System for Ordering Small and Large Numbers in Monkeys and Humans. *Psychological Science*, *17*(5), 401–406. doi: 10.1111/j.1467-9280.2006.01719.x

Cantlon, J. F., & Brannon, E. M. (2007). How much does number matter to a monkey (Macaca mulatta)? *Journal of Experimental Psychology: Animal Behavior Processes*, *33*(1), 32–41. doi: 10.1037/0097-7403.33.1.32

Cantlon, J. F., Brannon, E. M., Carter, E. J., & Pelphrey, K. A. (2006). Functional Imaging of Numerical Processing in Adults and 4-y-Old Children. *PLOS Biology*, *4*(5), e125. doi: 10.1371/journal.pbio.0040125

Cantlon, J. F., Platt, M. L., & Brannon, E. M. (2009). Beyond the number domain. *Trends in Cognitive Sciences*, *13*(2), 83–91. doi: 10.1016/j.tics.2008.11.007

Cappelletti, M., Butterworth, B., & Kopelman, M. (2006). The Understanding of Quantifiers in Semantic Dementia: A Single-Case Study. *Neurocase*, *12*(3), 136–145. doi: 10.1080/13554790600598782

Carey, S. (2001). Cognitive Foundations of Arithmetic: Evolution and Ontogenisis. *Mind & Language*, *16*(1), 37–55. doi: 10.1111/1468-0017.00155

Carey, S. (2009). Where Our Number Concepts Come From. *The journal of philosophy*, *106*(4), 220–254.

Carey, S., & Barner, D. (2019). Ontogenetic Origins of Human Integer Representations. *Trends in Cognitive Sciences*, *23*(10), 823–835. doi: 10.1016/j.tics.2019.07.004

Carey, S., Shusterman, A., Haward, P., & Distefano, R. (2017). Do analog number representations underlie the meanings of young children's verbal numerals?

Cognition, *168*, 243–255. doi: 10.1016/j.cognition.2017.06.022

Casasanto, D., & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition*, *106*(2), 579–593. doi: 10.1016/j.cognition.2007.03.004

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130. doi: 10.3758/s13428-013-0365-7

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using Nonnaive Participants Can Reduce Effect Sizes. *Psychological Science*, *26*(7), 1131–1139. doi: 10.1177/0956797615585115

Chassy, P., & Grodd, W. (2012). Comparison of Quantities: Core and Format-Dependent Regions as Revealed by fMRI. *Cerebral Cortex*, *22*(6), 1420–1430. doi: 10.1093/cercor/bhr219

Chen, Q., & Li, J. (2014). Association between individual differences in non-symbolic number acuity and math performance: A meta-analysis. *Acta Psychologica*, *148*, 163–172. doi: 10.1016/j.actpsy.2014.01.016

Cheng, D., Zhou, A., Yu, X., Chen, C., Jia, J., & Zhou, X. (2013). Quantifier processing can be dissociated from numerical processing: Evidence from semantic dementia patients. *Neuropsychologia*, *51*(11), 2172–2183. doi: 10.1016/j.neuropsychologia.2013.07.003

Cipolotti, L., & Butterworth, B. (1995). Toward a Multiroute Model of Number Processing: Impaired Number Transcoding With Preserved Calculation Skills. *Journal of Experimental Psychology: General*, *124*(4), 375–390.

Cipora, K., Soltanlou, M., Reips, U.-D., & Nuerk, H.-C. (2019). The SNARC and MARC effects measured online: Large-scale assessment methods in flexible cognitive effects. *Behavior Research Methods*, *51*, 1676–1692. doi: 10.3758/s13428-019-01213-5

Cipora, K., Willmes, K., Szwarc, A., & Nuerk, H.-C. (2018). Norms and Validation of the Online and Paper-and-Pencil Versions of the Abbreviated Math Anxiety Scale (AMAS) For Polish Adolescents and Adults. *Journal of Numerical Cognition*, *3*(3), 667–693. doi: 10.5964/jnc.v3i3.121

Clark, R., & Grossman, M. (2007). Number sense and quantifier interpretation. *Topoi*, *26*(1), 51–62.

Clementz, B. A., Barber, S. K., & Dzau, J. R. (2002). Knowledge of Stimulus Repetition Affects the Magnitude and Spatial Distribution of Low-Frequency Event-Related Brain Potentials. *Audiology and Neurotology*, *7*(5), 303–314. doi: 10.1159/000064444

Cohen, D. J. (2009). Integers do not automatically activate their quantity representation. *Psychonomic Bulletin & Review*, *16*(2), 332–336. doi: 10.3758/PBR.16.2.332

Cohen, D. J., Warren, E., & Blanc-Goldhammer, D. (2013). Cross-format physical similarity effects and their implications for the numerical cogni-

tion architecture. *Cognitive Psychology*, *66*(4), 355–379. doi: 10.1016/j.cogpsych.2013.03.001

Cohen Kadosh, R., Bahrami, B., Walsh, V., Butterworth, B., Popescu, T., & Price, C. J. (2011). Specialization in the Human Brain: The Case of Numbers. *Frontiers in Human Neuroscience*, *5*. doi: 10.3389/fnhum.2011.00062

Cohen Kadosh, R., Cohen Kadosh, K., Linden, D. E. J., Gevers, W., Berger, A., & Henik, A. (2007). The Brain Locus of Interaction between Number and Size: A Combined Functional Magnetic Resonance Imaging and Event-related Potential Study. *Journal of Cognitive Neuroscience*, *19*(6), 957–970. doi: 10.1162/jocn.2007.19.6.957

Cohen Kadosh, R., & Henik, A. (2006). A common representation for semantic and physical properties: A cognitive-anatomical approach. *Experimental Psychology*, *53*(2), 87–94. doi: 10.1027/1618-3169.53.2.87

Cohen Kadosh, R., Henik, A., & Rubinsten, O. (2008). Are Arabic and verbal numbers processed in different ways? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1377–1391. doi: 10.1037/a0013413

Cohen Kadosh, R., Lammertyn, J., & Izard, V. (2008). Are numbers special? An overview of chronometric, neuroimaging, developmental and comparative studies of magnitude representation. *Progress in Neurobiology*, *84*(2), 132–147. doi: 10.1016/j.pneurobio.2007.11.001

Cohen Kadosh, R., & Walsh, V. (2009). Numerical representation in the parietal lobes: Abstract or not abstract? *Behavioral and Brain Sciences*, *32*(3-4), 313. doi: 10.1017/S0140525X09990938

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi: 10.1126/science.aac4716

Cordes, S., Gelman, R., Gallistel, C. R., & Whalen, J. (2001). Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin & Review*, *8*(4), 698–707. doi: 10.3758/BF03196206

Coventry, K. R., Cangelosi, A., Newstead, S., Bacon, A., & Rajapakse, R. (2005). Grounding natural language quantifiers in visual attention. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 27, pp. 506–511).

Coventry, K. R., Cangelosi, A., Newstead, S. E., & Bugmann, D. (2010). Talking about quantities in space: Vague quantifiers, context and similarity. *Language and Cognition*, *2*(2), 221–241. doi: 10.1515/langcog.2010.009

Crollen, V., Castronovo, J., & Seron, X. (2011). Under- and over-estimation: A bi-directional mapping process between symbolic and non-symbolic representations of number? *Experimental Psychology*, *58*(1), 39–49. doi: 10.1027/1618-3169/a000064

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating

Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLOS ONE*, *8*(3), e57410. doi: 10.1371/journal.pone.0057410

Dadon, G., & Henik, A. (2017). Adjustment of control in the numerical Stroop task. *Memory & Cognition*, *45*(6), 891–902. doi: 10.3758/s13421-017-0703-6

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., & Halgren, E. (2000). Dynamic Statistical Parametric Mapping: Combining fMRI and MEG for High-Resolution Imaging of Cortical Activity. *Neuron*, *26*(1), 55–67. doi: 10.1016/S0896-6273(00)81138-1

de Hevia, M.-D., & Spelke, E. S. (2009). Spontaneous mapping of number and space in adults and young children. *Cognition*, *110*(2), 198–207. doi: 10.1016/j.cognition.2008.11.003

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. doi: 10.3758/s13428-014-0458-y

de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods*, *48*, 1–12. doi: 10.3758/s13428-015-0567-2

Defever, E., Sasanguie, D., Gebuis, T., & Reynvoet, B. (2011). Children's representation of symbolic and nonsymbolic magnitude examined with the priming paradigm. *Journal of Experimental Child Psychology*, *109*(2), 174–186. doi: 10.1016/j.jecp.2011.01.002

Defever, E., Sasanguie, D., Vandewaetere, M., & Reynvoet, B. (2012). What can the same–different task tell us about the development of magnitude representations? *Acta Psychologica*, *140*(1), 35–42. doi: 10.1016/j.actpsy.2012.02.005

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, *44*(1), 1–42. doi: 10.1016/0010-0277(92)90049-N

Dehaene, S. (1997). *The number sense. How the mind creates mathematics*. New York: Oxford University Press. doi: 10.2307/2589308

Dehaene, S. (2007). Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation. In P. Haggard, Y. Rosetti, & M. Kawato (Eds.), *Sensorimotor Foundations of Higher Cognition* (pp. 527–574). New York: Oxford University Press.

Dehaene, S., & Akhavein, R. (1995). Attention, Automaticity, and Levels of Representation in Number Processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 314–326. doi: 10.1037/0278-7393.21.2.314

Dehaene, S., & Changeux, J.-P. (1993). Development of Elementary Numerical Abilities: A Neuronal Model. *Journal of Cognitive Neuroscience*, *5*(4), 390–407. doi: 10.1162/jocn.1993.5.4.390

Dehaene, S., & Cohen, L. (2007). Cultural Recycling of Cortical Maps. *Neuron*,

56(2), 384–398. doi: 10.1016/j.neuron.2007.10.004

Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(3), 626–641. doi: 10.1037/0096-1523.16.3.626

Dehaene, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science*, *320*(5880), 1217–1220. doi: 10.1126/science.1156540

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., ... Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature*, *395*(6702), 597–600. doi: 10.1038/26967

Del Prato, P., & Pylkkänen, L. (2014). MEG evidence for conceptual combination but not numeral quantification in the left anterior temporal lobe during language production. *Frontiers in Psychology*, *5*. doi: 10.3389/fpsyg.2014.00524

Deschamps, I., Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2015). The processing of polar quantifiers, and numerosity perception. *Cognition*, *143*, 115–128. doi: 10.1016/j.cognition.2015.06.006

De Smedt, B., Noël, M.-P., Gilmore, C., & Ansari, D. (2013). How do symbolic and non-symbolic numerical magnitude processing skills relate to individual differences in children's mathematical skills? A review of evidence from brain and behavior. *Trends in Neuroscience and Education*, *2*(2), 48–55. doi: 10.1016/j.tine.2013.06.001

Dickey, J. M., & Lientz, B. P. (1970). The Weighted Likelihood Ratio, Sharp Hypotheses about Chances, the Order of a Markov Chain. *The Annals of Mathematical Statistics*, *41*(1), 214–226. doi: 10.1214/aoms/1177697203

Dolscheid, S., Winter, C., Ostrowski, L., & Penke, M. (2017). The many ways quantifiers count: Children's quantifier comprehension and cardinal number knowledge are not exclusively related. *Cognitive Development*, *44*, 21–31. doi: 10.1016/j.cogdev.2017.08.004

Dormal, V., Seron, X., & Pesenti, M. (2006). Numerosity-duration interference: A Stroop experiment. *Acta Psychologica*, *121*(2), 109–124. doi: 10.1016/j.actpsy.2005.06.003

Droit-Volet, S. (2010). Speeding up a master clock common to time, number and length? *Behavioural Processes*, *85*(2), 126–134. doi: 10.1016/j.beproc.2010.06.017

Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 488–499. doi: 10.3758/BF03195595

Eddelbuettel, D., & Balamuta, J. J. (2018). Extending R with C++: A Brief Introduction to Rcpp. *The American Statistician*, *72*(1), 28–36. doi: 10.1080/00031305.2017.1375990

Eger, E. (2016). Neuronal foundations of human numerical representations. In M. Cappelletti & W. Fias (Eds.), *Progress in Brain Research* (Vol. 227, pp. 1–27). Elsevier. doi: 10.1016/bs.pbr.2016.04.015

Eger, E., Michel, V., Thirion, B., Amadon, A., Dehaene, S., & Kleinschmidt, A. (2009). Deciphering cortical number coding from human brain activity patterns. *Current Biology*, *19*(19), 1608–1615.

Eger, E., Pinel, P., Dehaene, S., & Kleinschmidt, A. (2015). Spatially Invariant Coding of Numerical Information in Functionally Defined Subregions of Human Parietal Cortex. *Cerebral Cortex*, *25*(5), 1319–1329. doi: 10.1093/cercor/bht323

Égré, P. (2017). Vague Judgment: A Probabilistic Account. *Synthese*, *194*(10), 3837–3865. doi: 10.1007/s11229-016-1092-2

Égré, P., & Barberousse, A. (2014). Borel on the Heap. *Erkenntnis*, *79*(5), 1043–1079. doi: 10.1007/s10670-013-9596-3

Elze, T., & Tanner, T. G. (2012). Temporal Properties of Liquid Crystal Displays: Implications for Vision Science Experiments. *PLOS ONE*, *7*(9), e44048. doi: 10.1371/journal.pone.0044048

Epps, P., Bowern, C., Hansen, C. A., Hill, J. H., & Zentz, J. (2012). On numeral complexity in hunter-gatherer languages. *Linguistic Typology*, *16*(1), 41–109. doi: 10.1515/lity-2012-0002

Eriksson, K., & Lindskog, M. (2017). Encoding of numerical information in memory: Magnitude or nominal? *Journal of Numerical Cognition*. doi: http://dx.doi.org/10.23668/psycharchives.1440

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. doi: 10.1007/s11192-011-0494-7

Faulkenberry, T. J., Cruise, A., Lavro, D., & Shaki, S. (2016). Response trajectories capture the continuous dynamics of the size congruity effect. *Acta Psychologica*, *163*, 114–123. doi: 10.1016/j.actpsy.2015.11.010

Fazio, L. K., Bailey, D. H., Thompson, C. A., & Siegler, R. S. (2014). Relations of different types of numerical magnitude representations to each other and to mathematics achievement. *Journal of Experimental Child Psychology*, *123*, 53–72. doi: 10.1016/j.jecp.2014.01.013

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194. doi: 10.1152/jn.00032.2010

Feigenson, L. (2007). The equality of quantity. *Trends in Cognitive Sciences*, *11*(5), 185–187. doi: 10.1016/j.tics.2007.01.006

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. doi: 10.1016/j.tics.2004.05.002

Ferguson, A. M., Maloney, E. A., Fugelsang, J., & Risko, E. F. (2015). On the relation between math and spatial ability: The case of math anxiety. *Learning and Individual Differences*, *39*, 1–12. doi: 10.1016/j.lindif.2015

.02.007

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120–128. doi: 10.1037/a0024445

Ferrigno, S., Hughes, K. D., & Cantlon, J. F. (2016). Precocious quantitative cognition in monkeys. *Psychonomic Bulletin & Review*, *23*(1), 141–147. doi: 10.3758/s13423-015-0893-5

Ferrigno, S., Jara-Ettinger, J., Piantadosi, S. T., & Cantlon, J. F. (2017). Universal and uniquely human factors in spontaneous number perception. *Nature Communications*, *8*(1), 1–10. doi: 10.1038/ncomms13968

Fias, W., Lammertyn, J., Reynvoet, B., Dupont, P., & a Orban, G. (2003). Parietal representation of symbolic and nonsymbolic magnitude. *Journal of cognitive neuroscience*, *15*(1), 47–56. doi: 10.1162/089892903321107819

Flick, G., Oseki, Y., Kaczmarek, A. R., Al Kaabi, M., Marantz, A., & Pylkkänen, L. (2018). Building words and phrases in the left temporal lobe. *Cortex*, *106*, 213–236. doi: 10.1016/j.cortex.2018.06.004

Flick, G., & Pylkkänen, L. (2018). Isolating syntax in natural language: MEG evidence for an early contribution of left posterior temporal cortex. *bioRxiv*, 439158. doi: 10.1101/439158

Foltz, G. S., Poltrock, S. E., & Potts, G. R. (1984). Mental comparison of size and magnitude: Size congruity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(3), 442–453. doi: 10.1037/0278-7393.10.3.442

Ford, N., & Reynolds, M. G. (2016). Do Arabic numerals activate magnitude automatically? Evidence from the psychological refractory period paradigm. *Psychonomic Bulletin & Review*, *23*(5), 1528–1533. doi: 10.3758/s13423-016-1020-y

Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, *37*(2), 413–420.

Frazier, L., Clifton, C., & Stolterfoht, B. (2008). Scale structure: Processing minimum standard and maximum standard scalar adjectives. *Cognition*, *106*(1), 299–324. doi: 10.1016/J.COGNITION.2007.02.004

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, *6*(2), 78–84. doi: 10.1016/S1364-6613(00)01839-8

Friederici, A. D., Meyer, M., & von Cramon, D. Y. (2000). Auditory Language Comprehension: An Event-Related fMRI Study on the Processing of Syntactic and Lexical Information. *Brain and Language*, *74*(2), 289–300. doi: 10.1006/brln.2000.2313

Fults, S. (2011). Vagueness and scales. In P. Égré & N. Klinedinst (Eds.), *Vagueness and Language Use* (pp. 25–50). Palgrave Macmillan UK.

Gabay, S., Leibovich, T., Henik, A., & Gronau, N. (2013). Size before numbers:

Conceptual size primes numerical value. *Cognition*, *129*(1), 18–23. doi: 10.1016/j.cognition.2013.06.001

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, *44*(1), 43–74. doi: 10.1016/0010-0277(92)90050-R

Gallistel, C. R., & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences*, *4*(2), 59–65. doi: 10.1016/S1364-6613(99)01424-2

Ganor-Stern, D., & Tzelgov, J. (2008). Across-notation automatic numerical processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(2), 430–437. doi: 10.1037/0278-7393.34.2.430

Gebuis, T., Cohen Kadosh, R., & Gevers, W. (2016). Sensory-integration system rather than approximate number system underlies numerosity processing: A critical review. *Acta Psychologica*, *171*, 17–35. doi: 10.1016/j.actpsy.2016.09.003

Gelman, A. (2018). The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It. *Personality and Social Psychology Bulletin*, *44*(1), 16–23. doi: 10.1177/0146167217729162

Geyer, L. H. (1977). Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics*, *22*(5), 487–490. doi: 10.3758/BF03199515

Gibson, E., Jara-Ettinger, J., Levy, R., & Piantadosi, S. (2017). The Use of a Computer Display Exaggerates the Connection Between Education and Approximate Number Ability in Remote Populations. *Open Mind*, *2*(1), 37–46. doi: 10.1162/opmi_a_00016

Glasser, M. F., & Van Essen, D. C. (2011). Mapping human cortical areas in vivo based on myelin content as revealed by T1- and T2-weighted MRI. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(32), 11597–11616. doi: 10.1523/JNEUROSCI.2180-11.2011

Gleibs, I. H. (2017). Are all "research fields" equal? Rethinking practice for the use of data from crowdsourcing market places. *Behavior Research Methods*, *49*(4), 1333–1342. doi: 10.3758/s13428-016-0789-y

Gliksman, Y., Itamar, S., Leibovich, T., Melman, Y., & Henik, A. (2016). Automaticity of Conceptual Magnitude. *Scientific reports*, *6*, 21446. doi: 10.1038/srep21446

Goffin, C., Sokolowski, H. M., Slipenkyj, M., & Ansari, D. (2019). Does writing handedness affect neural representation of symbolic number? An fMRI adaptation study. *Cortex*, *121*, 27–43. doi: 10.1016/j.cortex.2019.07.017

Gökaydin, D., Brugger, P., & Loetscher, T. (2018). Sequential Effects in SNARC. *Scientific Reports*, *8*(1), 10996. doi: 10.1038/s41598-018-29337-2

Gordon, P. (2004). Numerical Cognition Without Words: Evidence from Amazonia. *Science*, *306*(5695), 496–499. doi: 10.1126/science.1094492

Gosling, S. D., & Mason, W. (2015). Internet Research in Psychology. *Annual Review of Psychology*, *66*(1), 877–902. doi: 10.1146/annurev-psych-010814-015321

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires. *American Psychologist*, *59*(2), 93–104. doi: 10.1037/0003-066X.59.2.93

Graff, F. D. (2000). Shifting Sands: An Interest-Relative Theory of Vagueness. *Philosophical Topics*, *28*(1), 45–81.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. *arXiv:1802.06893 [cs]*.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14–23. doi: 10.1016/j.tics.2005.11.006

Gunderson, E. A., Spaepen, E., & Levine, S. C. (2015). Approximate number word knowledge before the cardinal principle. *Journal of Experimental Child Psychology*, *130*, 35–55. doi: 10.1016/j.jecp.2014.09.008

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun - An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods*, *47*, 930–944. doi: 10.3758/s13428-014-0529-0

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: Most versus more than half. *Natural Language Semantics*, *17*(1), 63–98. doi: 10.1007/s11050-008-9039-x

Hagoort, P. (2003). Interplay between Syntax and Semantics during Sentence Comprehension: ERP Effects of Combining Syntactic and Semantic Violations. *Journal of Cognitive Neuroscience*, *15*(6), 883–899. doi: 10.1162/089892903322370807

Hagoort, P. (2017). The core and beyond in the language-ready brain. *Neuroscience & Biobehavioral Reviews*, *81*, 194–204. doi: 10.1016/j.neubiorev.2017.01.048

Hagoort, P., & Brown, C. M. (1999). Gender Electrified: ERP Evidence on the Syntactic Nature of Gender Processing. *Journal of Psycholinguistic Research*, *28*(6), 715–728. doi: 10.1023/A:1023277213129

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the "Number Sense": The Approximate Number System in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental psychology*, *44*(5), 1457–65. doi: 10.1037/a0012682

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D. Q., & Germine, L. (2012). Number sense across the lifespan as revealed by a massive Internet-based sample. *Proceedings of the National Academy of Sciences*, *109*(28), 11116–11120. doi: 10.1073/pnas.1200196109

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665–668. doi: 10.1038/nature07246

Halberda, J., Taing, L., & Lidz, J. (2008). The Development of "Most" Comprehension and Its Potential Dependence on Counting Ability in Preschool-

ers. *Language Learning and Development*, *4*(2), 99–121. doi: 10.1080/15475440801922099

Halpern, C., Glosser, G., Clark, R., Gee, J., Moore, P., Dennis, K., ... Grossman, M. (2004). Dissociation of numbers and objects in corticobasal degeneration and semantic dementia. *Neurology*, *62*(7), 1163–1169. doi: 10.1212/01.wnl.0000118209.95423.96

Halpern, C., McMillan, C., Moore, P., Dennis, K., & Grossman, M. (2003). Calculation impairment in neurodegenerative diseases. *Journal of the Neurological Sciences*, *208*(1), 31–38. doi: 10.1016/S0022-510X(02)00416-1

Hämäläinen, M. S., & Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: Minimum norm estimates. *Medical & Biological Engineering & Computing*, *32*(1), 35–42. doi: 10.1007/BF02512476

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1), 335–346. doi: 10.1016/0167-2789(90)90087-6

Harrell Jr, F. E. (2020). *Hmisc: Harrell miscellaneous*.

Harvey, B. M., Ferri, S., & Orban, G. A. (2017). Comparing Parietal Quantity-Processing Mechanisms between Humans and Macaques. *Trends in Cognitive Sciences*, *21*(10), 779–793. doi: 10.1016/j.tics.2017.07.002

Hauk, O., Coutout, C., Holden, A., & Chen, Y. (2012). The time-course of single-word reading: Evidence from fast behavioral and brain responses. *NeuroImage*, *60*(2), 1462–1477. doi: 10.1016/j.neuroimage.2012.01.061

Hauk, O., Davis, M. H., & Pulvermüller, F. (2008). Modulation of brain activity by multiple lexical and word form variables in visual word recognition: A parametric fMRI study. *NeuroImage*, *42*(3), 1185–1195. doi: 10.1016/j.neuroimage.2008.05.054

Hauk, O., & Pulvermüller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, *115*(5), 1090–1103. doi: 10.1016/j.clinph.2003.12.020

He, L., Zhou, K., Zhou, T., He, S., & Chen, L. (2015). Topology-defined units in numerosity perception. *Proceedings of the National Academy of Sciences*, *112*(41), E5647-E5655. doi: 10.1073/pnas.1512408112

Heim, S., Amunts, K., Drai, D., Eickhoff, S., Hautvast, S., & Grodzinsky, Y. (2012). The Language–Number Interface in the Brain: A Complex Parametric Study of Quantifiers and Quantities. *Frontiers in Evolutionary Neuroscience*, *4*. doi: 10.3389/fnevo.2012.00004

Heim, S., McMillan, C. T., Clark, R., Baehr, L., Ternes, K., Olm, C., ... Grossman, M. (2016). How the brain learns how few are "many": An fMRI study of the flexibility of quantifier semantics. *NeuroImage*, *125*, 45–52. doi: 10.1016/j.neuroimage.2015.10.035

Heim, S., McMillan, C. T., Clark, R., Golob, S., Min, N. E., Olm, C., ... Grossman, M. (2015). If so many are “few,” how few are “many”? *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.00441

Henik, A., Gliksman, Y., Kallai, A., & Leibovich, T. (2017). Size Perception and

the Foundation of Numerical Processing. *Current Directions in Psychological Science*, *26*(1), 45–51. doi: 10.1177/0963721416671323

Henik, A., & Tzelgov, J. (1982). Is three greater than five: The relation between physical and semantic size in comparison tasks. *Memory & Cognition*, *10*(4), 389–395. doi: 10.3758/BF03202431

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2019). *Lab.js: A free, open, online study builder* (Preprint). PsyArXiv. doi: 10.31234/osf.io/fqr49

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. doi: 10.1017/S0140525X0999152X

Herrera, A., & Macizo, P. (2008). Cross-notational semantic priming between symbolic and nonsymbolic numerosity. *Quarterly Journal of Experimental Psychology*, *61*(10), 1538–1552. doi: 10.1080/17470210701595530

Hilbig, B. E. (2016). Reaction time effects in lab- versus Web-based research: Experimental evidence. *Behavior Research Methods*, *48*(4), 1718–1724. doi: 10.3758/s13428-015-0678-9

Hinrichs, J. V., Yurko, D. S., & Hu, J.-m. (1981). Two-digit number comparison: Use of place information. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(4), 890–901. doi: 10.1037/0096-1523.7.4.890

Holloway, I. D., Battista, C., Vogel, S. E., & Ansari, D. (2012). Semantic and Perceptual Processing of Number Symbols: Evidence from a Cross-linguistic fMRI Adaptation Study. *Journal of Cognitive Neuroscience*, *25*(3), 388–400. doi: 10.1162/jocn_a_00323

Holyoak, K. J., & Glass, A. L. (1978). Recognition confusions among quantifiers. *Journal of verbal learning and verbal behavior*, *17*(3), 249–264.

Huber, S., Nuerk, H.-C., Reips, U.-D., & Soltanlou, M. (2017). Individual differences influence two-digit number processing, but not their analog magnitude processing: A large-scale online study. *Psychological Research*. doi: 10.1007/s00426-017-0964-5

Hultén, A., Schoffelen, J.-M., Uddén, J., Lam, N. H. L., & Hagoort, P. (2019). How the brain makes sense beyond the processing of single words – An MEG study. *NeuroImage*, *186*, 586–594. doi: 10.1016/j.neuroimage.2018.11.035

Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2006). Syntactic and Semantic Modulation of Neural Activity during Auditory Sentence Comprehension. *Journal of Cognitive Neuroscience*, *18*(4), 665–679. doi: 10.1162/jocn.2006.18.4.665

Hurewitz, F., Gelman, R., & Schnitzer, B. (2006). Sometimes area counts more than number. *Proceedings of the National Academy of Sciences*, *103*(51), 19599–19604. doi: 10.1073/pnas.0609485103

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. doi: 10.1207/s15473341lld0202_1

Hutchison, J. E., Ansari, D., Zheng, S., Jesus, S. D., & Lyons, I. M. (2019). The relation between subitizable symbolic and non-symbolic number processing over the kindergarten school year. *Developmental Science*, *0*(0), e12884. doi: 10.1111/desc.12884

Hyde, D., & Raffman, D. (2018). Sorites Paradox. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2018 ed.). Metaphysics Research Lab, Stanford University.

Hyde, D. C., Khanum, S., & Spelke, E. S. (2014). Brief non-symbolic, approximate number practice enhances subsequent exact symbolic arithmetic in children. *Cognition*, *131*(1), 92–107. doi: 10.1016/j.cognition.2013.12.007

Im, H. Y., Zhong, S.-h., & Halberda, J. (2016). Grouping by proximity and the visual impression of approximate number in random dot arrays. *Vision Research*, *126*, 291–307. doi: 10.1016/j.visres.2015.08.013

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology*, *19*(5), 640–648. doi: 10.1097/EDE.0b013e31818131e7

Irvine, A. (2018). *The Unity game engine for a large-cohort, developmental study.* Behavioral Science Online conference, London.

Ito, Y., & Hatta, T. (2003). Semantic processing of Arabic, Kanji, and Kana numbers: Evidence from interference in physical and numerical size judgments. *Memory & Cognition*, *31*(3), 360–368. doi: 10.3758/BF03194394

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, *106*(3), 1221–1247. doi: 10.1016/j.cognition.2007.06.004

Izard, V., Dehaene-Lambertz, G., & Dehaene, S. (2008). Distinct Cerebral Pathways for Object Identity and Number in Human Infants. *PLOS Biology*, *6*(2), e11. doi: 10.1371/journal.pbio.0060011

Jacob, S. N., & Nieder, A. (2009a). Notation-Independent Representation of Fractions in the Human Parietal Cortex. *Journal of Neuroscience*, *29*(14), 4652–4657. doi: 10.1523/JNEUROSCI.0651-09.2009

Jacob, S. N., & Nieder, A. (2009b). Tuning to non-symbolic proportions in the human frontoparietal cortex. *European Journal of Neuroscience*, *30*(7), 1432–1442. doi: 10.1111/j.1460-9568.2009.06932.x

Jacob, S. N., Vallentin, D., & Nieder, A. (2012). Relating magnitudes: The brain's code for proportions. *Trends in Cognitive Sciences*, *16*(3), 157–166. doi: 10.1016/J.TICS.2012.02.002

Jeffreys, S. H. (1998). *The Theory of Probability* (Third Edition ed.). Oxford, New York: Oxford University Press.

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes*, *15*(2), 159–201. doi: 10.1080/016909600386084

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, *57*(2), 129–191. doi: 10.1016/0010-0277(94)00659-9

Katsos, N., Cummins, C., Ezeizabarrena, M.-J., Gavarró, A., Kraljević, J. K.,

Hrzica, G., ... Noveck, I. (2016). Cross-linguistic patterns in the acquisition of quantifiers. *Proceedings of the National Academy of Sciences*, *113*(33), 9244–9249. doi: 10.1073/pnas.1601341113

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The Discrimination of Visual Number. *The American Journal of Psychology*, *62*(4), 498–525. doi: 10.2307/1418556

Kaufmann, L., Koppelstaetter, F., Delazer, M., Siedentopf, C., Rhomberg, P., Golaszewski, S., ... Ischebeck, A. (2005). Neural correlates of distance and congruity effects in a numerical Stroop task: An event-related fMRI study. *NeuroImage*, *25*(3), 888–898. doi: 10.1016/j.neuroimage.2004.12.041

Keenan, E. L. (2012). The Quantifier Questionnaire. In E. L. Keenan & D. Paperno (Eds.), *Handbook of Quantifiers in Natural Language* (pp. 1–20). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-007-2681-9_1

Keller, L., & Libertus, M. (2015). Inhibitory control may not explain the link between approximation and math abilities in kindergarteners from middle class families. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.00685

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, *30*(1), 1–45. doi: 10.1007/s10988-006-9008-0

Kennedy, C., & McNally, L. (2005). Scale Structure, Degree Modification, and the Semantics of Gradable Predicates. *Language*, *81*(2), 345–381. doi: 10.1353/lan.2005.0071

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. doi: 10.3758/BRM.42.3.643

Kim, S., & Pylkkänen, L. (2019). Composition of event concepts: Evidence for distinct roles for the left and right anterior temporal lobes. *Brain and Language*, *188*, 18–27. doi: 10.1016/j.bandl.2018.11.003

Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and Philosophy*, *4*(1), 1–45. doi: 10.1007/BF00351812

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., ... Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. doi: 10.1177/2515245918810225

Knowlton, T., Pietroski, P., Halberda, J., & Lidz, J. (2020). *The mental representation of universal quantifiers.*

Kochari, A. (2019). Conducting Web-Based Experiments for Numerical Cognition Research. *Journal of Cognition*, *2*(1), 39. doi: 10.5334/joc.85

Kochari, A., & Flecken, M. (2019). Lexical prediction in language comprehension: A replication study of grammatical gender effects in Dutch. *Language, Cognition and Neuroscience*, *34*(2), 239–253. doi: 10.1080/23273798.2018.1524500

Kochari, A., Lewis, A., Schoffelen, J.-M., & Schriefers, H. (2020). Semantic

and syntactic composition of minimal adjective-noun phrases in Dutch: An MEG study. *bioRxiv*, 2020.03.14.991802. doi: 10.1101/2020.03.14.991802

Kochari, A., & Ostarek, M. (2018). Introducing a replication-first rule for Ph.D. projects. *The Behavioral and Brain Sciences*, *41*, e138. doi: 10.1017/S0140525X18000730

Koechlin, E., Naccache, L., Block, E., & Dehaene, S. (1999). Primed numbers: Exploring the modularity of numerical representations with masked and unmasked semantic priming. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(6), 1882–1905. doi: 10.1037/0096-1523.25.6.1882

Kojouharova, P., & Krajcsi, A. (2018). The Indo-Arabic distance effect originates in the response statistics of the task. *Psychological Research*. doi: 10.1007/s00426-018-1052-1

Konkle, T., & Oliva, A. (2012). A familiar-size Stroop effect: Real-world size is an automatic property of object representation. *Journal of experimental psychology. Human perception and performance*, *38*(3), 561–9. doi: 10.1037/a0028294

Koss, S., Clark, R., Vesely, L., Weinstein, J., Powers, C., Richmond, L., ... Grossman, M. (2010). Numerosity impairment in corticobasal syndrome. *Neuropsychology*, *24*(4), 476–492. doi: 10.1037/a0018755

Kounios, J., & Holcomb, P. J. (1994). Concreteness effects in semantic processing: ERP evidence supporting dual-coding theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 804.

Krajcsi, A., Lengyel, G., & Kojouharova, P. (2016). The Source of the Symbolic Numerical Distance and Size Effects. *Frontiers in Psychology*, *7*, 1795. doi: 10.3389/fpsyg.2016.01795

Krajcsi, A., Lengyel, G., & Kojouharova, P. (2018). Symbolic Number Comparison Is Not Processed by the Analog Number System: Different Symbolic and Non-symbolic Numerical Distance and Size Effects. *Frontiers in Psychology*, *9*. doi: 10.3389/fpsyg.2018.00124

Krause, F., Bekkering, H., Pratt, J., & Lindemann, O. (2017). Interaction between numbers and size during visual search. *Psychological Research*, *81*(3), 664–677. doi: 10.1007/s00426-016-0771-4

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. doi: 10.3389/neuro.06.004.2008

Krueger, L. E. (1982). Single judgments of numerosity. *Perception & Psychophysics*, *31*(2), 175–182. doi: 10.3758/BF03206218

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. doi: 10.1146/annurev.psych.093008.131123

Kutter, E. F., Bostroem, J., Elger, C. E., Mormann, F., & Nieder, A. (2018).

Single Neurons in the Human Brain Encode Numbers. *Neuron*, *100*(3), 753-761.e4. doi: 10.1016/j.neuron.2018.08.036

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. doi: 10.18637/jss.v082.i13

Landy, D., Silbert, N., & Goldin, A. (2013). Estimating Large Numbers. *Cognitive Science*, *37*(5), 775–799. doi: 10.1111/cogs.12028

Lassiter, D. (2015). Adjectival Modification and Gradation. In S. Lappin & C. Fox (Eds.), *The Handbook of Contemporary Semantic Theory* (pp. 141–167). Wiley-Blackwell. doi: 10.1002/9781118882139.ch5

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positiveform adjectives. *Proceedings of SALT 23*, 587610. doi: 10.3765/salt.v23i0.2658

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. doi: 10.1038/nrn2532

Lawrence, M. A. (2016). *Ez: Easy Analysis and Visualization of Factorial Experiments*.

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: 10.1016/j.cognition.2006.10.005

Leibovich, T., & Ansari, D. (2016). The symbol-grounding problem in numerical cognition: A review of theory, evidence, and outstanding questions. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *70*(1), 12–23. doi: 10.1037/cep0000070

Leibovich, T., Diesendruck, L., Rubinsten, O., & Henik, A. (2013). The importance of being relevant: Modulation of magnitude representations. *Frontiers in Psychology*, *4*. doi: 10.3389/fpsyg.2013.00369

Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From 'sense of number' to 'sense of magnitude': The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences*, *40*, E164. doi: 10.1017/S0140525X16000960

Libertus, M. E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental Science*, *14*(6), 1292–1300. doi: 10.1111/j.1467-7687.2011.01080.x

Lidz, J., Pietroski, P., Halberda, J., & Hunter, T. (2011). Interface transparency and the psychosemantics of most. *Natural Language Semantics*, *19*(3), 227–256. doi: 10.1007/s11050-010-9062-6

Lin, C.-Y., & Göbel, S. M. (2019). Arabic digits and spoken number words: Timing modulates the cross-modal numerical distance effect. *Quarterly Journal of Experimental Psychology*, *72*(11), 2632–2646. doi: 10.1177/1747021819854444

Lindelø v, J. K. (2020). *Reaction time distributions: An interactive overview.*

https://lindeloev.shinyapps.io/shiny-rt/.

Liu, A. S., Schunn, C. D., Fiez, J. A., & Libertus, M. E. (2015). Symbolic Integration, Not Symbolic Estrangement, For Double-Digit Numbers. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Liu, R., Schunn, C. D., Fiez, J. A., & Libertus, M. E. (2018). The integration between nonsymbolic and symbolic numbers: Evidence from an EEG study. *Brain and Behavior*, *8*(4). doi: 10.1002/brb3.938

Lourenco, S. F. (2015). On the relation between numerical and non-numerical magnitudes: Evidence for a general magnitude system. In D. C. Geary, D. B. Berch, & K. M. Koepke (Eds.), *Evolutionary origins and early development of number processing* (Vol. 1, pp. 145–174). Elsevier. doi: 10.1016/B978-0-12-420133-0.00006-5

Lourenco, S. F., Ayzenberg, V., & Lyu, J. (2016). A general magnitude system in human adults: Evidence from a subliminal priming paradigm. *Cortex*, *81*, 93–103. doi: 10.1016/j.cortex.2016.04.013

Lyle, H., Wylie, J., & Morsanyi, K. (2019). *Cross-cultural differences in children's mathematical development: Investigating the home numeracy environment.* Ottawa ON, Canada.

Lyons, I. M., Ansari, D., & Beilock, S. L. (2012). Symbolic estrangement: Evidence against a strong association between numerical symbols and the quantities they represent. *Journal of Experimental Psychology: General*, *141*(4), 635–641. doi: 10.1037/a0027248

Lyons, I. M., Ansari, D., & Beilock, S. L. (2015). Qualitatively different coding of symbolic and nonsymbolic numbers in the human brain. *Human Brain Mapping*, *36*(2), 475–488. doi: 10.1002/hbm.22641

Lyons, I. M., & Beilock, S. L. (2018). Characterizing the neural coding of symbolic quantities. *NeuroImage*, *178*, 503–518. doi: 10.1016/j.neuroimage.2018.05.062

Makuuchi, M., Bahlmann, J., Anwander, A., & Friederici, A. D. (2009). Segregating the core computational faculty of human language from working memory. *Proceedings of the National Academy of Sciences*, *106*(20), 8362–8367. doi: 10.1073/pnas.0810928106

Marcus, D., Harwell, J., Olsen, T., Hodge, M., Glasser, M., Prior, F., ... Van Essen, D. (2011). Informatics and Data Mining Tools and Strategies for the Human Connectome Project. *Frontiers in Neuroinformatics*, *5*. doi: 10.3389/fninf.2011.00004

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. doi: 10.1016/j.jneumeth.2007.03.024

Matchin, W., Hammerly, C., & Lau, E. (2017). The role of the IFG and pSTS in syntactic prediction: Evidence from a parametric study of hierarchical structure in fMRI. *Cortex*, *88*, 106–123. doi: 10.1016/j.cortex.2016.12.010

Matchin, W., & Hickok, G. (2020). The Cortical Organization of Syntax. *Cerebral Cortex*, *30*(3), 1481–1498. doi: 10.1093/cercor/bhz180

Matejko, A. A., & Ansari, D. (2016). Trajectories of Symbolic and Nonsymbolic Magnitude Processing in the First Year of Formal Schooling. *PLOS ONE*, *11*(3), e0149863. doi: 10.1371/journal.pone.0149863

Mathur, M., & Reichling, D. (2018). Open-source software for mouse-tracking in Qualtrics to measure category competition. *Behavior Research Methods*, *51*, 1987–1997. doi: 10.3758/s13428-019-01258-6

Mazoyer, B., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., ... Mehler, J. (1993). The cortical representation of speech. *Journal of Cognitive Neuroscience*, *5*(4), 467–479.

Mazoyer, B., Zago, L., Jobard, G., Crivello, F., Joliot, M., Perchey, G., ... Tzourio-Mazoyer, N. (2014). Gaussian Mixture Modeling of Hemispheric Lateralization for Language in a Large Sample of Healthy Individuals Balanced for Handedness. *PLOS ONE*, *9*(6), e101165. doi: 10.1371/journal.pone.0101165

Mazzocco, M. M. M., Feigenson, L., & Halberda, J. (2011). Impaired Acuity of the Approximate Number System Underlies Mathematical Learning Disability (Dyscalculia). *Child Development*, *82*(4), 1224–1237. doi: 10.1111/j.1467-8624.2011.01608.x

McCloskey, M. (1992). Cognitive mechanisms in numerical processing: Evidence from acquired dyscalculia. *Cognition*, *44*(1), 107–157. doi: 10.1016/0010-0277(92)90052-J

McMillan, C. T., Clark, R., Moore, P., Devita, C., & Grossman, M. (2005). Neural basis for generalized quantifier comprehension. *Neuropsychologia*, *43*(12), 1729–1737. doi: 10.1016/j.neuropsychologia.2005.02.012

McMillan, C. T., Clark, R., Moore, P., & Grossman, M. (2006). Quantifier comprehension in corticobasal degeneration. *Brain and Cognition*, *62*(3), 250–260. doi: 10.1016/j.bandc.2006.06.005

Meade, G., Grainger, J., & Holcomb, P. J. (2019). Task modulates ERP effects of orthographic neighborhood for pseudowords but not words. *Neuropsychologia*, *129*, 385–396. doi: 10.1016/j.neuropsychologia.2019.02.014

Merritt, D. J., Casasanto, D., & Brannon, E. M. (2010). Do monkeys think in metaphors? Representations of space and time in monkeys and humans. *Cognition*, *117*(2), 191–202. doi: 10.1016/j.cognition.2010.08.011

Merten, K., & Nieder, A. (2008). Compressed Scaling of Abstract Numerosity Representations in Adult Humans and Monkeys. *Journal of Cognitive Neuroscience*, *21*(2), 333–346. doi: 10.1162/jocn.2008.21032

Miller, K. F., & Stigler, J. W. (1987). Counting in Chinese: Cultural variation in a basic cognitive skill. *Cognitive Development*, *2*(3), 279–305. doi: 10.1016/S0885-2014(87)90091-8

Möhring, W., Ramsook, K. A., Hirsh-Pasek, K., Golinkoff, R. M., & Newcombe, N. S. (2016). *Where music meets space: Children's sensitivity to pitch*

intervals is related to their mental spatial transformation skills (Vol. 151; Tech. Rep.). doi: 10.1016/j.cognition.2016.02.016

Molinaro, N., Barber, H. A., & Carreiras, M. (2011). Grammatical agreement processing in reading: ERP findings and future directions. *Cortex*, *47*(8), 908–930. doi: 10.1016/j.cortex.2011.02.019

Morgan, B., Gross, R. G., Clark, R., Dreyfuss, M., Boller, A., Camp, E., ... Grossman, M. (2011). Some is not enough: Quantifier comprehension in corticobasal syndrome and behavioral variant frontotemporal dementia. *Neuropsychologia*, *49*(13), 3532–3541. doi: 10.1016/j.neuropsychologia.2011.09.005

Moss, H. E., McCormick, S. F., & Tyler, L. K. (1997). The Time Course of Activation of Semantic Information during Spoken Word Recognition. *Language and Cognitive Processes*, *12*(5-6), 695–732. doi: 10.1080/016909697386664

Moxey, L. M., & Sanford, A. J. (1993). Prior expectation and the interpretation of natural language quantifiers. *European Journal of Cognitive Psychology*, *5*(1), 73–91.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*(5109), 1519–1520. doi: 10.1038/2151519a0

Mundy, E., & Gilmore, C. K. (2009). Children's mapping between symbolic and nonsymbolic representations of number. *Journal of Experimental Child Psychology*, *103*(4), 490–502. doi: 10.1016/j.jecp.2009.02.003

Mussolin, C., Nys, J., Leybaert, J., & Content, A. (2012). Relationships between approximate number system acuity and early symbolic number abilities. *Trends in Neuroscience and Education*, *1*(1), 21–31. doi: 10.1016/j.tine.2012.09.003

Naccache, L., & Dehaene, S. (2001a). The Priming Method: Imaging Unconscious Repetition Priming Reveals an Abstract Representation of Number in the Parietal Lobes. *Cerebral Cortex*, *11*(10), 966–974. doi: 10.1093/cercor/11.10.966

Naccache, L., & Dehaene, S. (2001b). Unconscious semantic priming extends to novel unseen stimuli. *Cognition*, *80*(3), 215–229. doi: 10.1016/S0010-0277(00)00139-6

Nalborczyk, L., Batailler, C., Lœ venbruck, H., Vilain, A., & Bürkner, P.-C. (2019). An Introduction to Bayesian Multilevel Models Using brms: A Case Study of Gender Effects on Vowel Variability in Standard Indonesian. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1225–1242. doi: 10.1044/2018_JSLHR-S-18-0006

Nation, K. (2018). *Online large-scale studies with children out of the lab: The promise and the challenge.* Behavioral Science Online conference, London.

Neath, I., Earle, A., Hallett, D., & Surprenant, A. M. (2011). Response time accuracy in Apple Macintosh computers. *Behavior Research Methods*, *43*(2), 353. doi: 10.3758/s13428-011-0069-9

Necka, E. A., Cacioppo, S., Norman, G. J., & Cacioppo, J. T. (2016). Mea-

suring the Prevalence of Problematic Respondent Behaviors among MTurk, Campus, and Community Participants. *PLOS ONE*, *11*(6), e0157732. doi: 10.1371/journal.pone.0157732

Neufeld, C., Kramer, S. E., Lapinskaya, N., Heffner, C. C., Malko, A., & Lau, E. F. (2016). The Electrophysiology of Basic Phrase Building. *PLOS ONE*, *11*(10), e0158446. doi: 10.1371/journal.pone.0158446

Newstead, S. E., & Coventry, K. R. (2000). The role of context and functionality in the interpretation of quantifiers. *European Journal of Cognitive Psychology*, *12*(2), 243–259. doi: 10.1080/095414400382145

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas—Part II. *Language and Linguistics Compass*, *10*(11), 591–613. doi: 10.1111/lnc3.12207

Nieder, A. (2016). The neuronal code for number. *Nature Reviews Neuroscience*, *17*(6), 366–382. doi: 10.1038/nrn.2016.40

Nieder, A., & Dehaene, S. (2009). Representation of Number in the Brain. *Annual Review of Neuroscience*, *32*(1), 185–208. doi: 10.1146/annurev.neuro.051508.135550

Nieder, A., Freedman, D. J., & Miller, E. K. (2002). Representation of the Quantity of Visual Items in the Primate Prefrontal Cortex. *Science*, *297*(5587), 1708–1711. doi: 10.1126/science.1072493

Nieder, A., & Miller, E. K. (2003). Coding of Cognitive Magnitude: Compressed Scaling of Numerical Information in the Primate Prefrontal Cortex. *Neuron*, *37*(1), 149–157. doi: 10.1016/S0896-6273(02)01144-3

Nieder, A., & Miller, E. K. (2004). A parieto-frontal network for visual numerical information in the monkey. *Proceedings of the National Academy of Sciences*, *101*(19), 7457–7462. doi: 10.1073/pnas.0402239101

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., ... Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468. doi: 10.7554/eLife.33468

Nolte, G. (2003). The magnetic lead field theorem in the quasi-static approximation and its use for magnetoencephalography forward calculation in realistic volume conductors. *Physics in Medicine and Biology*, *48*(22), 3637–3652. doi: 10.1088/0031-9155/48/22/002

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. doi: 10.1073/pnas.1708274114

Notebaert, K., Nelis, S., & Reynvoet, B. (2010). The Magnitude Representation of Small and Large Symbolic Numbers in the Left and Right Hemisphere: An Event-related fMRI Study. *Journal of Cognitive Neuroscience*, *23*(3), 622–630. doi: 10.1162/jocn.2010.21445

Notebaert, K., Pesenti, M., & Reynvoet, B. (2010). The neural origin of the priming distance effect: Distance-dependent recovery of parietal activation

using symbolic magnitudes. *Human Brain Mapping*, *31*(5), 669–677. doi: 10.1002/hbm.20896

Núñez, R. E. (2017). Is There Really an Evolved Capacity for Number? *Trends in Cognitive Sciences*, *21*(6), 409–424. doi: 10.1016/j.tics.2017.03.005

Odic, D., Le Corre, M., & Halberda, J. (2015). Children's mappings between number words and the approximate number system. *Cognition*, *138*, 102–121. doi: 10.1016/j.cognition.2015.01.008

Odic, D., Pietroski, P., Hunter, T., Lidz, J., & Halberda, J. (2013). Young children's understanding of "more" and discrimination of number and surface area. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 451–461. doi: 10.1037/a0028874

Oliveri, M., Vicario, C. M., Salerno, S., Koch, G., Turriziani, P., Mangano, R., ... Caltagirone, C. (2008). Perceiving numbers alters time perception. *Neuroscience Letters*, *438*(3), 308–311. doi: 10.1016/j.neulet.2008.04.051

Olm, C. A., McMillan, C. T., Spotorno, N., Clark, R., & Grossman, M. (2014). The relative contributions of frontal and parietal cortex for generalized quantifier comprehension. *Frontiers in Human Neuroscience*, *8*. doi: 10.3389/fnhum.2014.00610

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience*, *2011*, 1.

Pagin, P. (2010). Vagueness and Central Gaps. In R. Dietz & S. Moruzzi (Eds.), *Cuts and Clouds: Vagueness, its Nature, & its Logic* (pp. 254–272). Oxford University Press. doi: 10.1093/acprof:oso/9780199570386.003.0015

Paivio, A. (1975). Perceptual comparisons through the mind's eye. *Memory & Cognition*, *3*(6), 635–647. doi: 10.3758/BF03198229

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. doi: 10.1016/j.jbef.2017.12.004

Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, *108*(6), 2522–2527.

Pansky, A., & Algom, D. (1999). Stroop and Garner effects in comparative judgment of numerals: The role of attention. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(1), 39.

Pansky, A., & Algom, D. (2002). Comparative judgment of numerosity and numerical magnitude: Attention preempts automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(2), 259–274. doi: 10.1037/0278-7393.28.2.259

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, *23*(3), 184–188. doi: 10.1177/0963721414531598

Paperno, D., & Keenan, E. L. (Eds.). (2017). *Handbook of Quantifiers in Natural*

Language: Volume II. Springer International Publishing. doi: 10.1007/978-3-319-44330-0

Park, J., & Brannon, E. M. (2013). Training the Approximate Number System Improves Math Proficiency. *Psychological Science*, *24*(10), 2013–2019. doi: 10.1177/0956797613482944

Partee, B. (1995). Lexical semantics and compositionality. *An invitation to cognitive science: Language*, *1*, 311–360.

Partee, B. (2004). Many Quantifiers. In *Compositionality in Formal Semantics: Selected Papers* (pp. 241–258). Wiley-Blackwell.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. doi: 10.1016/j.jesp.2017.01.006

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... Lindelø v, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. doi: 10.3758/s13428-018-01193-y

Peters, S., & Westerståh l, D. (2006). *Quantifiers in Language and Logic*. Oxford, New York: Oxford University Press.

Pezzelle, S., Bernardi, R., & Piazza, M. (2018). Probing the mental representation of quantifiers. *Cognition*, *181*, 117–126. doi: 10.1016/j.cognition.2018.08.009

Piantadosi, S. T., Jara-Ettinger, J., & Gibson, E. (2014). Children's learning of number words in an indigenous farming-foraging group. *Developmental Science*, *17*(4), 553–563. doi: 10.1111/desc.12078

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: 10.1016/j.cognition.2011.11.005

Piazza, M., & Eger, E. (2016). Neural foundations and functional specificity of number representations. *Neuropsychologia*, *83*, 257–273. doi: 10.1016/j.neuropsychologia.2015.09.025

Piazza, M., Facoetti, A., Trussardi, A. N., Berteletti, I., Conte, S., Lucangeli, D., ... Zorzi, M. (2010). Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition*, *116*(1), 33–41. doi: 10.1016/j.cognition.2010.03.012

Piazza, M., Fumarola, A., Chinello, A., & Melcher, D. (2011). Subitizing reflects visuo-spatial object individuation capacity. *Cognition*, *121*(1), 147–153. doi: 10.1016/j.cognition.2011.05.007

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, *44*(3), 547–555. doi: 10.1016/j.neuron.2004.10.014

Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A Magnitude Code Common to Numerosities and Number Symbols in Human Intraparietal

Cortex. *Neuron*, *53*(2), 293–305. doi: 10.1016/J.NEURON.2006.11.022

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and Approximate Arithmetic in an Amazonian Indigene Group. *Science*, *306*(5695), 499–503. doi: 10.1126/science.1102085

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The Meaning of 'Most': Semantics, Numerosity and Psychology. *Mind & Language*, *24*(5), 554–585. doi: 10.1111/j.1468-0017.2009.01374.x

Pinel, P., Piazza, M., Le Bihan, D., & Dehaene, S. (2004a). Distributed and Overlapping Cerebral Representations of Number, Size, and Luminance during Comparative Judgments. *Neuron*, *41*(6), 983–993. doi: 10.1016/S0896-6273(04)00107-2

Pinel, P., Piazza, M., Le Bihan, D., & Dehaene, S. (2004b). Distributed and overlapping cerebral representations of number, size, and luminance during comparative judgments. *Neuron*, *41*(6), 983–993. doi: 10.1016/S0896-6273(04)00107-2

Pitt, B., & Casasanto, D. (2016). Reading experience shapes the mental timeline but not the mental number line. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2753–2758). Austin, TX: Cognitive Science Society.

Plant, R. R., & Turner, G. (2009). Millisecond precision psychological research in a world of commodity computers: New hardware, new problems? *Behavior Research Methods*, *41*(3), 598–614. doi: 10.3758/BRM.41.3.598

Poortman, E. B., & Pylkkänen, L. (2016). Adjective conjunction as a window into the LATL's contribution to conceptual combination. *Brain and Language*, *160*, 50–60. doi: 10.1016/j.bandl.2016.07.006

Price, A. R., Bonner, M. F., Peelle, J. E., & Grossman, M. (2015). Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *Journal of Neuroscience*, *35*(7), 3276–3284.

Price, A. R., Peelle, J. E., Bonner, M. F., Grossman, M., & Hamilton, R. H. (2016). Causal Evidence for a Mechanism of Semantic Integration in the Angular Gyrus as Revealed by High-Definition Transcranial Direct Current Stimulation. *Journal of Neuroscience*, *36*(13), 3829–3838. doi: 10.1523/JNEUROSCI.3120-15.2016

PrincetonUniversity. (2010). *Princeton University "About WordNet"*. https://wordnet.princeton.edu/.

Proctor, R. W., & Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin*, *132*(3), 416–442. doi: 10.1037/0033-2909.132.3.416

Pylkkänen, L. (2016). Chapter 50 - Composition of Complex Meaning: Interdisciplinary Perspectives on the Left Anterior Temporal Lobe. In G. Hickok & S. L. Small (Eds.), *Neurobiology of Language* (pp. 621–631). San Diego: Academic Press. doi: 10.1016/B978-0-12-407794-2.00050-X

Pylkkänen, L. (2019). The neural basis of combinatory syntax and semantics.

Science, *366*(6461), 62–66. doi: 10.1126/science.aax0050

Pylkkänen, L., Bemis, D. K., & Blanco Elorrieta, E. (2014). Building phrases in language production: An MEG study of simple composition. *Cognition*, *133*(2), 371–384. doi: 10.1016/j.cognition.2014.07.001

Pylkkänen, L., & Brennan, J. R. (2019). Composition: The neurobiology of syntactic and semantic structure building. In *The Cognitive Neurosciences* (Sixth ed.). MIT Press.

Qing, C., & Franke, M. (2014). Meaning and use of gradable adjectives: Formal modeling meets empirical data. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55. doi: 10.1038/nrn.2016.150

Ramotowska, S., Steinert-Threlkeld, S., Leendert, V. M., & Szymanik, J. (2020). Most, but not more than half, is proportion-dependent and sensitive to individual differences. In *Proceedings of Sinn und Bedeutung 2020*.

Register, J., Mollica, F., & Piantadosi, S. T. (2020). *Semantic verification is flexible and sensitive to context.*

Reike, D., & Schwarz, W. (2017). Exploring the origin of the number-size congruency effect: Sensitivity or response bias? *Attention, Perception, & Psychophysics*, *79*(2), 383–388. doi: 10.3758/s13414-016-1267-4

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. *Behavior Research Methods*, *47*(2), 309–327. doi: 10.3758/s13428-014-0471-1

Reimers, S., & Stewart, N. (2016). Auditory presentation and synchronization in Adobe Flash and HTML5/JavaScript Web experiments. *Behavior Research Methods*, *48*(3), 897–908. doi: 10.3758/s13428-016-0758-5

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does Subitizing Reflect Numerical Estimation? *Psychological Science*, *19*(6), 607–614. doi: 10.1111/j.1467-9280.2008.02130.x

Reynvoet, B., & Brysbaert, M. (1999). Single-digit and two-digit Arabic numerals address the same semantic number line. *Cognition*, *72*(2), 191–201. doi: 10.1016/S0010-0277(99)00048-7

Reynvoet, B., Brysbaert, M., & Fias, W. (2002). Semantic priming in number naming. *The Quarterly Journal of Experimental Psychology Section A*, *55*(4), 1127–1139. doi: 10.1080/02724980244000116

Reynvoet, B., De Smedt, B., & Van den Bussche, E. (2009). Children's representation of symbolic magnitude: The development of the priming distance effect. *Journal of Experimental Child Psychology*, *103*(4), 480–489. doi:

10.1016/j.jecp.2009.01.007

Reynvoet, B., & Sasanguie, D. (2016). The Symbol Grounding Problem Revisited: A Thorough Evaluation of the ANS Mapping Account and the Proposal of an Alternative Account Based on Symbol–Symbol Associations. *Frontiers in Psychology*, *7*. doi: 10.3389/fpsyg.2016.01581

Risko, E. F., Maloney, E. A., & Fugelsang, J. A. (2013). Paying attention to attention: Evidence for an attentional contribution to the size congruity effect. *Attention, Perception, & Psychophysics*, *75*(6), 1137–1147. doi: 10.3758/s13414-013-0477-2

Rogalsky, C., & Hickok, G. (2009). Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. *Cerebral Cortex*, *19*(4), 786–796. doi: 10.1093/cercor/bhn126

Roggeman, C., Verguts, T., & Fias, W. (2007). Priming reveals differential coding of symbolic and non-symbolic quantities. *Cognition*, *105*(2), 380–394. doi: 10.1016/j.cognition.2006.10.004

Roitman, J. D., Brannon, E. M., Andrews, J. R., & Platt, M. L. (2007). Nonverbal representation of time and number in adults. *Acta Psychologica*, *124*(3), 296–318. doi: 10.1016/j.actpsy.2006.03.008

Rousselet, G. A., & Wilcox, R. R. (2019). Reaction times and other skewed distributions: Problems with the mean and the median. *bioRxiv*, 383935. doi: 10.1101/383935

Routh, D. A. (1994). On Representations of Quantifiers. *Journal of Semantics*, *11*(3), 199–214. doi: 10.1093/jos/11.3.199

Rubinsten, O., & Henik, A. (2002). Is an ant larger than a lion? *Acta Psychologica*, *111*(1), 141–154. doi: 10.1016/S0001-6918(02)00047-1

Rubio-Fernandez, P., Terrasa, H. A., Shukla, V., & Jara-Ettinger, J. (2019). Contrastive inferences are sensitive to informativity expectations, adjective semantics and visual salience. *PsyArXiv*. doi: 10.31234/osf.io/mr4ah

Santens, S., & Verguts, T. (2011). The size congruity effect: Is bigger always more? *Cognition*, *118*(1), 94–110. doi: 10.1016/j.cognition.2010.10.014

Sasanguie, D., Defever, E., Van den Bussche, E., & Reynvoet, B. (2011). The reliability of and the relation between non-symbolic numerical distance effects in comparison, same-different judgments and priming. *Acta Psychologica*, *136*(1), 73–80. doi: 10.1016/j.actpsy.2010.10.004

Sasanguie, D., De Smedt, B., & Reynvoet, B. (2017). Evidence for distinct magnitude systems for symbolic and non-symbolic number. *Psychological Research*, *81*(1), 231–242. doi: 10.1007/s00426-015-0734-1

Sassenhagen, J., & Alday, P. M. (2016). A common misapplication of statistical inference: Nuisance control with null-hypothesis significance tests. *Brain and Language*, *162*, 42–45. doi: 10.1016/j.bandl.2016.08.001

Sassoon, G. W. (2013). A Typology of Multidimensional Adjectives. *Journal of Semantics*, *30*(3), 335–380. doi: 10.1093/jos/ffs012

Scarf, D., Hayne, H., & Colombo, M. (2011). Pigeons on Par with Primates

in Numerical Competence. *Science*, *334*(6063), 1664–1664. doi: 10.1126/science.1213357

Schell, M., Zaccarella, E., & Friederici, A. D. (2017). Differential cortical contribution of syntax and semantics: An fMRI study on two-word phrasal processing. *Cortex*, *96*, 105–120. doi: 10.1016/j.cortex.2017.09.002

Schlotterbeck, F., Ramotowska, S., van Maanen, L., & Szymanik, J. (2020). *Is the monotonicity effect due to covert negation or pragmatic bias?*

Schmidt, S. (2009). Shall we Really do it Again? The Powerful Concept of Replication is Neglected in the Social Sciences. *Review of General Psychology*, *13*(2), 90–100. doi: 10.1037/a0015108

Schneider, M., Beeres, K., Coban, L., Merz, S., Schmidt, S. S., Stricker, J., & Smedt, B. D. (2017). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: A meta-analysis. *Developmental Science*, *20*(3), e12372. doi: 10.1111/desc.12372

Schurz, M., Sturm, D., Richlan, F., Kronbichler, M., Ladurner, G., & Wimmer, H. (2010). A dual-route perspective on brain activation in response to visual words: Evidence for a length by lexicality interaction in the visual word form area (VWFA). *NeuroImage*, *49*(3), 2649–2661. doi: 10.1016/j.neuroimage.2009.10.082

Schwarz, W., & Heinze, H.-J. (1998). On the interaction of numerical and size information in digit comparison: A behavioral and event-related potential study. *Neuropsychologia*, *36*(11), 1167–1179. doi: 10.1016/S0028-3932(98)00001-3

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147. doi: 10.1016/S0010-0277(99)00025-6

Sekuler, R., Rubin, E., & Armstrong, R. (1971). Processing numerical information: A choice time analysis. *Journal of Experimental Psychology*, *90*(1), 75–80. doi: 10.1037/h0031366

Semmelmann, K., Hönekopp, A., & Weigelt, S. (2017). Looking Tasks Online: Utilizing Webcams to Collect Video Data from Home. *Frontiers in Psychology*, *8*. doi: 10.3389/fpsyg.2017.01582

Semmelmann, K., & Weigelt, S. (2017). Online psychophysics: Reaction time effects in cognitive experiments. *Behavior Research Methods*, *49*(4), 1241–1260. doi: 10.3758/s13428-016-0783-4

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods*, *50*(2), 451–465. doi: 10.3758/s13428-017-0913-7

Shaki, S., Fischer, M. H., & Petrusic, W. M. (2009). Reading habits for both words and numbers contribute to the SNARC effect. *Psychonomic Bulletin & Review*, *16*(2), 328–331. doi: 10.3758/PBR.16.2.328

Shikhare, S., Heim, S., Klein, E., Huber, S., & Willmes, K. (2015). Processing of Numerical and Proportional Quantifiers. *Cognitive Science*, *39*(7), 1504–

1536. doi: 10.1111/cogs.12219

Siegel, M. (1976). *Capturing the Adjective* (PhD Thesis). University of Massachusets Amherst.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. doi: 10.1177/0956797611417632

Smets, K., Gebuis, T., & Reynvoet, B. (2013). Comparing the neural distance effect derived from the non–symbolic comparison and the same–different task. *Frontiers in Human Neuroscience*, *7*. doi: 10.3389/fnhum.2013.00028

Smets, K., Moors, P., & Reynvoet, B. (2016). Effects of Presentation Type and Visual Control in Numerosity Discrimination: Implications for Number Processing? *Frontiers in Psychology*, *7*. doi: 10.3389/fpsyg.2016.00066

Sokolowski, H. M., & Ansari, D. (2016). Symbolic and Nonsymbolic Representation of Number in the Human Parietal Cortex: A Review of the State-of-the-Art, Outstanding Questions and Future Directions. In A. Henik (Ed.), *Continuous Issues in Numerical Cognition* (pp. 326–353). San Diego: Academic Press. doi: 10.1016/B978-0-12-801637-4.00015-9

Sokolowski, H. M., Fias, W., Bosah Ononye, C., & Ansari, D. (2017). Are numbers grounded in a general magnitude processing system? A functional neuroimaging meta-analysis. *Neuropsychologia*. doi: 10.1016/j.neuropsychologia.2017.01.019

Sokolowski, H. M., Fias, W., Mousa, A., & Ansari, D. (2017). Common and distinct brain regions in both parietal and frontal cortex support symbolic and nonsymbolic number processing in humans: A functional neuroimaging meta-analysis. *NeuroImage*, *146*, 376–394. doi: 10.1016/J.NEUROIMAGE.2016.10.028

Solt, S. (2011). Vagueness in Quantity: Two Case Studies from a Linguistic Perspective. In P. Cintula, C. Fermüller, L. Godo, & P. Hájek (Eds.), *Understanding Vagueness. Logical, Philosophical and Linguistic Perspectives* (Vol. 36, pp. 157–174). London: College Publications.

Solt, S. (2015a). Q-Adjectives and the Semantics of Quantity. *Journal of Semantics*, *32*(2), 221–273. doi: 10.1093/jos/fft018

Solt, S. (2015b). Vagueness and Imprecision: Empirical Foundations. *Annual Review of Linguistics*, *1*(1), 107–127. doi: 10.1146/annurev-linguist-030514-125150

Solt, S. (2016). On measurement and quantification: The case of most and more than half. *Language*(92), 65–100.

Solt, S. (2019). Adjective Meaning and Scales. In C. Cummins & N. Katsos (Eds.), *The Oxford Handbook of Experimental Semantics and Pragmatics.* Oxford University Press. doi: 10.1093/oxfordhb/9780198791768.013.27

Spelke, E. S. (2011). Natural Number and Natural Geometry. In S. Dehaene & E. M. Brannon (Eds.), *Space, Time and Number in the Brain* (pp. 287–317).

San Diego: Academic Press. doi: 10.1016/B978-0-12-385948-8.00018-9

Spotorno, N., McMillan, C. T., Powers, J. P., Clark, R., & Grossman, M. (2014). Counting or chunking? Mathematical and heuristic abilities in patients with corticobasal syndrome and posterior cortical atrophy. *Neuropsychologia*, *64*, 176–183. doi: 10.1016/j.neuropsychologia.2014.09.030

Starkey, P., & Cooper, R. G. (1995). The development of subitizing in young children. *British Journal of Developmental Psychology*, *13*(4), 399–420. doi: 10.1111/j.2044-835X.1995.tb00688.x

Steinert-Threlkeld, S., Munneke, G.-J., & Szymanik, J. (2015). Alternative representations in formal semantics: A case study of quantifiers. In T. Brochhagen, F. Roelofsen, & N. Theiler (Eds.), *Proceedings of the 20th Amsterdam Colloquium* (pp. 368–377).

Stevens, J. C., Mack, J. D., & Stevens, S. S. (1960). Growth of sensation on seven continua as measured by force of handgrip. *Journal of Experimental Psychology*, *59*(1), 60–67. doi: 10.1037/h0040746

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing Samples in Cognitive Science. *Trends in Cognitive Sciences*, *21*(10), 736–748. doi: 10.1016/j.tics.2017.06.007

Stewart, N., Ungemach, C., Harris, A. J. L., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making*, *10*(5), 479–491.

Stoet, G. (2017). PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. *Teaching of Psychology*, *44*(1), 24–31. doi: 10.1177/0098628316677643

Stolk, A., Todorovic, A., Schoffelen, J.-M., & Oostenveld, R. (2013). Online and offline tools for head movement compensation in MEG. *NeuroImage*, *68*, 39–48. doi: 10.1016/j.neuroimage.2012.11.047

Sullivan, J., Bale, A., & Barner, D. (2018). Most Preschoolers Don't Know Most. *Language Learning and Development*, *14*(4), 320–338. doi: 10.1080/15475441.2018.1489813

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), e2000797. doi: 10.1371/journal.pbio.2000797

Szucs, D., & Soltesz, F. (2008). The interaction of task-relevant and task-irrelevant stimulus features in the number/size congruency paradigm: An ERP study. *Brain research*, *1190*(1), 143–58. doi: 10.1016/j.brainres.2007.11.010

Szymanik, J. (2016a). Cognitive Processing of Quantifiers. In *Quantifiers and Cognition: Logical and Computational Perspectives* (pp. 51–83). Springer International Publishing.

Szymanik, J. (2016b). *Quantifiers and Cognition: Logical and Computational Perspectives*. Springer International Publishing. doi: 10.1007/978-3-319

-28749-2

Szymanik, J., & Zajenkowski, M. (2010). Comprehension of Simple Quantifiers: Empirical Evaluation of a Computational Model. *Cognitive Science*, *34*(3), 521–532. doi: 10.1111/j.1551-6709.2009.01078.x

Talmina, N., Kochari, A., & Szymanik, J. (2017). Quantifiers and verification strategies: Connecting the dots. In A. Cremens, T. van Gessel, & F. Roelofsen (Eds.), *Proceedings of the 21st Amsterdam Colloquium* (pp. 465–473).

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review*, *101*(1), 80–102. doi: 10.1037/0033-295X.101.1.80

Troiani, V., Clark, R., & Grossman, M. (2011). Impaired verbal comprehension of quantifiers in corticobasal syndrome. *Neuropsychology*, *25*(2), 159–165. doi: 10.1037/a0021448

Troiani, V., Peelle, J. E., Clark, R., & Grossman, M. (2009). Is it Logical to Count on Quantifiers? Dissociable Neural Networks Underlying Numerical and Logical Quantifiers. *Neuropsychologia*, *47*(1), 104–111. doi: 10.1016/j.neuropsychologia.2008.08.015

Tudusciuc, O., & Nieder, A. (2009). Contributions of primate prefrontal and posterior parietal cortices to length and numerosity representation. *Journal of Neurophysiology*, *101*(6), 2984–2994. doi: 10.1152/jn.90713.2008

Tzelgov, J., Meyer, J., & Henik, A. (1992). Automatic and intentional processing of numerical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(1), 166–179. doi: 10.1037/0278-7393.18.1.166

Vakharia, D., & Lease, M. (2015). Beyond Mechanical Turk: An Analysis of Paid Crowd Work Platforms. In *iConference, 2015.* Newport Beach, CA, USA.

Vallentin, D., & Nieder, A. (2008). Behavioral and prefrontal representation of spatial proportions in the monkey. *Current biology*, *18*(18), 1420–1425. doi: 10.1016/j.cub.2008.08.042

van Galen, M. S., & Reitsma, P. (2008). Developing access to number magnitude: A study of the SNARC effect in 7- to 9-year-olds. *Journal of Experimental Child Psychology*, *101*(2), 99–113. doi: 10.1016/j.jecp.2008.05.001

van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. doi: 10.1080/17470218.2013.850521

van Rooij, R. (2011a). Implicit versus explicit comparatives. In P. Egre & N. Klinedinst (Eds.), *Vagueness and language use* (pp. 51–22). Palgrave Macmillan UK.

van Rooij, R. (2011b). Vagueness and Linguistics. In *Vagueness: A Guide* (pp. 123–170). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-007-0375-9_6

Van Essen, D. C., Glasser, M. F., Dierker, D. L., Harwell, J., & Coalson, T. (2012). Parcellations and hemispheric asymmetries of human cerebral cortex

analyzed on surface-based atlases. *Cerebral Cortex*, *22*(10), 2241–2262. doi: 10.1093/cercor/bhr291

Van Opstal, F., Gevers, W., De Moor, W., & Verguts, T. (2008). Dissecting the symbolic distance effect: Comparison and priming effects in numerical and nonnumerical orders. *Psychonomic Bulletin & Review*, *15*(2), 419–425. doi: 10.3758/PBR.15.2.419

Van Opstal, F., & Verguts, T. (2011). The origins of the numerical distance effect: The same–different task. *Journal of Cognitive Psychology*, *23*(1), 112–120. doi: 10.1080/20445911.2011.466796

Van Opstal, F., & Verguts, T. (2013). Is there a generalized magnitude system in the brain? Behavioral, neuroimaging, and computational evidence. *Frontiers in Psychology*, *4*, 2011–2013. doi: 10.3389/fpsyg.2013.00435

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175. doi: 10.1016/j.jml.2018.07.004

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 147–161. doi: 10.1016/j.wocn.2018.07.008

Verguts, T., & Fias, W. (2004). Representation of Number in Animals and Humans: A Neural Model. *Journal of Cognitive Neuroscience*, *16*(9), 1493–1504. doi: 10.1162/0898929042568497

Verguts, T., Fias, W., & Stevens, M. (2005). A model of exact small-number representation. *Psychonomic Bulletin & Review*, *12*(1), 66–80. doi: 10.3758/BF03196349

Verguts, T., & Van Opstal, F. (2005). Dissociation of the distance effect and size effect in one-digit numbers. *Psychonomic Bulletin & Review*, *12*(5), 925–930. doi: 10.3758/BF03196787

Verguts, T., & Van Opstal, F. (2014). A delta-rule model of numerical and nonnumerical order processing. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1092–1102. doi: 10.1037/a0035114

Vogel, S. E., Goffin, C., Bohnenberger, J., Koschutnig, K., Reishofer, G., Grabner, R. H., & Ansari, D. (2017). The left intraparietal sulcus adapts to symbolic number in both the visual and auditory modalities: Evidence from fMRI. *NeuroImage*, *153*, 16–27. doi: 10.1016/j.neuroimage.2017.03.048

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189. doi: 10.1016/j.cogpsych.2009.12.001

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, *7*(6), 632–638. doi: 10.1177/1745691612463078

Wagner, J. B., & Johnson, S. C. (2011). An association between understanding cardinality and analog magnitude representations in preschoolers. *Cogni-*

tion, *119*(1), 10–22. doi: 10.1016/j.cognition.2010.11.014

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*(11), 483–488. doi: 10.1016/j.tics.2003.09.002

Walsh, V. (2015). A Theory of Magnitude: The Parts that Sum to Number. In R. Cohen Kadosh & A. Dowker (Eds.), *The Oxford Handbook of Numerical Cognition* (pp. 552–565). Oxford University Press.

Watson, D. G., Maylor, E. A., & Bruce, L. A. M. (2007). The role of eye movements in subitizing and counting. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1389–1399. doi: 10.1037/0096-1523.33.6.1389

Weber, S. J., & Cook, T. D. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin*, *77*(4), 273–295. doi: 10.1037/h0032351

Wei, W., Chen, C., Yang, T., Zhang, H., & Zhou, X. (2014). Dissociated neural correlates of quantity processing of quantifiers, numbers, and numerosities. *Human Brain Mapping*, *35*(2), 444–454. doi: 10.1002/hbm.22190

Westerlund, M., Kastner, I., Al Kaabi, M., & Pylkkänen, L. (2015). The LATL as locus of composition: MEG evidence from English and Arabic. *Brain and Language*, *141*, 124–134. doi: 10.1016/j.bandl.2014.12.003

Westerlund, M., & Pylkkänen, L. (2014). The role of the left anterior temporal lobe in semantic composition vs. semantic memory. *Neuropsychologia*, *57*, 59–70. doi: 10.1016/J.NEUROPSYCHOLOGIA.2014.03.001

Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal Counting in Humans: The Psychophysics of Number Representation:. *Psychological Science*, *10*(2). doi: 10.1111/1467-9280.00120

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Wickham, H., Hester, J., & Francois, R. (2018). *Readr: Read rectangular text data*.

Wilkey, E. D., & Ansari, D. (2019). Challenging the neurobiological link between number sense and symbolic numerical abilities. *Annals of the New York Academy of Sciences*. doi: 10.1111/nyas.14225

Wolter, L., Gorman, K. S., & Tanenhaus, M. K. (2011). Scalar reference, contrast and discourse: Separating effects of linguistic discourse from availability of the referent. *Journal of Memory and Language*, *65*(3), 299–317. doi: 10.1016/j.jml.2011.04.010

Wong, B., Bull, R., & Ansari, D. (2018). Magnitude processing of written number words is influenced by task, rather than notation. *Acta Psychologica*, *191*, 160–170. doi: 10.1016/j.actpsy.2018.09.010

Wong, B., & Szücs, D. (2013). Single-digit Arabic numbers do not automatically

activate magnitude representations in adults or in children: Evidence from the symbolic same–different task. *Acta Psychologica*, *144*(3), 488–498. doi: 10.1016/j.actpsy.2013.08.006

Woods, A. T., Velasco, C., Levitan, C. A., Wan, X., & Spence, C. (2015). Conducting perception research over the internet: A tutorial review. *PeerJ*, *3*, e1058. doi: 10.7717/peerj.1058

Wright, C. (1975). On the Coherence of Vague Predicates. *Synthese*, *30*(3/4), 325–365.

Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155–193. doi: 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: 10.1016/0010-0285(92)90008-P

Wynn, K. (1998). Psychological foundations of number: Numerical competence in human infants. *Trends in Cognitive Sciences*, *2*(8), 296–303. doi: 10.1016/S1364-6613(98)01203-0

Xie, Y. (2014). Knitr: A Comprehensive Tool for Reproducible Research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing Reproducible Research* (pp. 3–31). New York: Chapman and Hall/CRC. doi: 10.1201/9781315373461-1

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), B1-B11. doi: 10.1016/S0010-0277(99)00066-9

Xuan, B., Zhang, D., He, S., Chen, X., R., J. S., & F., M. (2007). Larger stimuli are judged to last longer. *Journal of Vision*, *7*(10), 2. doi: 10.1167/7.10.2

Yates, M. J., Loetscher, T., Nicholls, M. E. R., X., C., P., T., & R., M. (2012). A generalized magnitude system for space, time, and quantity? A cautionary note. *Journal of Vision*, *12*(7), 9–9. doi: 10.1167/12.7.9

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*, *87*, 128–143. doi: 10.1016/j.jml.2015.08.003

Zaccarella, E., & Friederici, A. D. (2015). Merge in the Human Brain: A Sub-Region Based Functional Investigation in the Left Pars Opercularis. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.01818

Zaccarella, E., Meyer, L., Makuuchi, M., & Friederici, A. D. (2017). Building by Syntax: The Neural Basis of Minimal Linguistic Structures. *Cerebral Cortex*, *27*(1), 411–421. doi: 10.1093/cercor/bhv234

Zajenkowski, M., Stył a, R., & Szymanik, J. (2011). A computational approach to quantifiers as an explanation for some language impairments in schizophrenia. *Journal of Communication Disorders*, *44*(6), 595–600. doi: 10.1016/j.jcomdis.2011.07.005

Zajenkowski, M., Szymanik, J., & Garraffa, M. (2014). Working Memory Mechanism in Proportional Quantifier Verification. *Journal of Psycholinguistic Research*, *43*(6), 839–853. doi: 10.1007/s10936-013-9281-3

Zhang, L., & Pylkkänen, L. (2015). The interplay of composition and concept specificity in the left anterior temporal lobe: An MEG study. *NeuroImage*, *111*, 228–240. doi: 10.1016/j.neuroimage.2015.02.028

Zhang, L., & Pylkkänen, L. (2018). Composing lexical versus functional adjectives: Evidence for uniformity in the left temporal lobe. *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-018-1469-y

Zhang, L., Xin, Z., Feng, T., Chen, Y., & Szűcs, D. (2017). Physical similarity or numerical representation counts in same–different, numerical comparison, physical comparison, and priming tasks? *Quarterly Journal of Experimental Psychology*, *71*(3), 670–687.

Ziegler, J., & Pylkkänen, L. (2016). Scalar adjectives and the temporal unfolding of semantic composition: An MEG investigation. *Neuropsychologia*, *89*, 161–171. doi: 10.1016/j.neuropsychologia.2016.06.010

Zorzi, M., Priftis, K., & Umiltà, C. (2002). Brain damage: Neglect disrupts the mental number line. *Nature*, *417*(6885), 138–139. doi: 10.1038/417138a

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*. doi: 10.1017/S0140525X17001972

Zwaan, R. A., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., & Zeelenberg, R. (2018). Participant Nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic Bulletin & Review*, *25*(5), 1968–1972. doi: 10.3758/s13423-017-1348-y

Samenvatting

In dit proefschrift worden onderzoeken gepresenteerd naar de cognitieve en neuronale processen die een rol spelen bij het begrijpen en produceren van schaalbare bijvoeglijke naamwoorden zoals 'groot', 'lang', 'hard', en 'zacht' en kwantoren zoals 'veel', 'weinig' en 'de meeste'. Het hoofdonderwerp van dit proefschrift (*hoofdstuk 2-4*) heeft betrekking op de potentiële relatie tussen enerzijds mechanismen voor de verwerking van schaalbare bijvoeglijke naamwoorden en kwantoren in natuurlijke taal (d.w.z. symbolische grootheden) en anderzijds verwerkingsmechanismen voor het inschatten en vergelijken van perceptueel gegeven (d.w.z. niet-symbolische) grootheden zoals hoeveelheid, lengte en duur op basis van perceptuele input. Schaalbare bijvoeglijke naamwoorden en in ieder geval sommige kwantoren in de natuurlijke taal kunnen worden beschouwd als verwijzingen naar niet-symbolische representaties van grootheden. De potentiële relatie tussen schaalbare bijvoeglijke naamwoorden en kwantoren enerzijds en perceptuele kwantiteiten anderzijds is onderzocht aan de hand van de vraag of deze op een vergelijkbare manier worden verwerkt als getalsymbolen (bijvoorbeeld Arabische cijfers zoals '3' en '5' en telwoorden zoals 'drie' en 'vijf'), een symbolische representatie van grootheid waarvan de interactie met niet-symbolische representaties van grootheid in het verleden al uitgebreid is onderzocht. In *hoofdstuk 5* wordt het proefschrift uitgebreid tot een onderzoek naar neuronale activiteit tijdens het combineren van schaalbare bijvoeglijke naamwoorden en zelfstandige naamwoorden.

Hoofdstuk 2 geeft een overzicht van nieuwe bevindingen en methodes van onderzoek naar de relatie tussen de verwerking van getalsymbolen en niet-symbolische grootheden door de hersenen. We bespreken bevindingen en methodes die mogelijk zouden kunnen worden gebruikt bij onderzoek naar de parallelle relatie tussen kwantoren en de verwerking van niet-symbolische grootheden. Ook bevat dit hoofdstuk een uitgebreide bespreking van de eigenschappen van verschillende klassen kwantoren in relatie tot de eigenschappen van mechanismen voor de verwerking van niet-symbolische grootheden. Daarnaast presenteren we een

verzameling onderzoeksrichtingen en specifieke vragen voor onderzoek naar de verwerking van kwantoren.

In *hoofdstuk 3* wordt ingegaan op aspecten van gegevensverzameling via het web voor onderzoek naar numerieke cognitie en worden resultaten gepresenteerd van twee replicatiestudies naar klassieke paradigma's binnen het onderzoek naar numerieke cognitie: het *aantal-grootte-congruentieparadigma* en *vergelijking met een bepaalde standaard*. Aan de hand van een van deze paradigma's, het aantal-grootte-congruentieparadigma, is vervolgens in *hoofdstuk 4* de verwerking van schaalbare bijvoeglijke naamwoorden onderzocht.

Hoofdstuk 4 bestaat uit een reeks van zes experimenten waarbij de interactie wordt onderzocht tussen de betekenis van telwoorden en schaalbare bijvoeglijke naamwoorden met niet-symbolische grootheden. Tijdens de kritische experimenten voor dit onderzoek werd de hypothese getoetst dat bij het oproepen van de betekenis van schaalbare bijvoeglijke naamwoorden het gegeneraliseerde representatiesysteem voor grootheden een onmisbare rol speelt.

De nadruk ligt in *hoofdstuk 5* op de hersenactiviteit die gepaard gaat met het maken van minimale combinaties van bijvoeglijke en zelfstandige naamwoorden. De betekenis van schaalbare bijvoeglijke naamwoorden is sterk afhankelijk van het zelfstandige naamwoord waarmee ze worden gecombineerd (bij 'grote stoel' tegenover 'groot huis', bijvoorbeeld, beschrijft het bijvoeglijke naamwoord 'groot' objecten die sterk van grootte verschillen), terwijl de betekenis van niet-gradeerbare bijvoeglijke naamwoorden (zoals 'dood', 'rechthoekig', 'houten', 'elektrisch') niet in dezelfde mate afhankelijk is van de betekenis van het zelfstandig naamwoord. Bij een eerder magneto-encefalografisch onderzoek werden verschillen in neuronale activiteit aangetoond tussen de verwerking van combinaties van schaalbare bijvoeglijke naamwoorden met zelfstandige naamwoorden en de verwerking van combinaties van nietgradeerbare bijvoeglijke naamwoorden met zelfstandige naamwoorden. Vermoedelijk zijn deze verschillen te verklaren vanuit het feit dat de betekenis van een schaalbaar bijvoeglijk naamwoord sterk afhankelijk is van de betekenis van het zelfstandig naamwoord, terwijl bij niet-gradeerbare bijvoeglijke naamwoorden geen sprake is van zo'n sterke contextafhankelijkheid. Ons onderzoek bouwde voort op deze bevindingen met als doel de robuustheid te bepalen van de waargenomen verschillen in neuronale activiteit.

Abstract

This thesis presents investigations of the cognitive and neuronal processes that take part in the comprehension and production of scalar adjectives such as 'large', 'long', 'loud', 'quiet' and quantifiers such as 'many', 'few', 'most'. The main topic of this thesis (*Chapters 2-4*) concerns the potential relationship between processing mechanisms for scalar adjectives and natural language quantifiers (i.e. symbolic magnitudes) on the one hand, and processing mechanisms for the estimation and comparison of perceptually given (i.e. nonsymbolic) magnitudes such as quantity, length, duration from perceptual input on other other hand. Scalar adjectives and at least some natural language quantifiers can be seen as references to nonsymbolic magnitude representations. The potential relation of scalar adjectives and quantifiers with perceptual quantities is investigated by asking whether they are processed in a similar way as number symbols (such as Arabic digits, e.g., '3', '5', and number words, e.g., 'three', 'five'), a symbolic magnitude representation whose interaction with nonsymbolic magnitude representations has already been a subject of extensive research in the past. In *Chapter 5* the scope of the thesis is expanded to an investigation of neuronal activity during composition of scalar adjectives and nouns.

Chapter 2 provides an overview of findings and methods from research into the relationship between number symbol and nonsymbolic magnitude processing by the brain. We review findings and methods that could be of potential use for research into the parallel relationship between quantifiers and nonsymbolic magnitude processing. Furthermore, this chapter presents an extended discussion on the properties of various quantifier classes in relation to the properties of nonsymbolic magnitude processing mechanisms. Importantly, we also provide a set of research directions and specific questions for the investigation of quantifier processing.

Chapter 3 reviews issues around web-based data collection for the purpose of numerical cognition research and presents results of two replication studies of classical paradigms in numerical cognition research: the *number size congruity*

paradigm and *comparison to a given standard*. One of these paradigms, the number size congruity paradigm, was subsequently used to investigate scalar adjective processing in *Chapter 4*.

Chapter 4 consists of a series of six experiments which explore the interaction of the meaning of number words and scalar adjectives with nonsymbolic magnitudes. The critical experiments of this study tested the hypothesis that retrieval of the meaning of scalar adjectives requires the involvement of the generalized magnitude representation system.

The focus of *Chapter 5* is on the brain activity accompanying the composition of minimal adjective-noun phrases. The meaning of scalar adjectives is highly dependent on the noun that they are combined with (e.g., in 'large chair' vs. 'large house', the adjective 'large' describes objects that are widely different in size), whereas the meaning of non-gradable adjectives (such as e.g., 'dead', 'rectangular', 'wooden', 'electric') is not dependent on the noun meaning to the same extent. A previous magnetoencephalography study reported differences in neuronal activity when processing adjective-noun phrases with scalar adjectives versus processing adjective-noun phrases with non-gradable adjectives. Presumably, these differences reflect the fact that the meaning of the adjective is highly dependent on on the noun meaning in the case of scalar adjectives, while such strong context dependency does not hold in the case of non-gradable adjectives. Our study followed up on these findings with the goal of determining the robustness of the observed differences in the neuronal activity.

Curriculum Vitae

Arnold Kochari was born 1990 in Qabala, Azerbaijan Soviet Socialist Republic, USSR. After finishing high school in Aktau, Kazakhstan, he moved to Prague, Czech Republic where in 2013 he completed a Bachelor's in Liberal Arts and Humanities at Charles University. During this period, he spent one semester as an exchange student at Amsterdam University College. Thereafter he received an Utrecht Excellence Scholarship to pursue a Research Master's programme in Linguistics at Utrecht University, from which he graduated cum laude in 2015. While at Utrecht University, he completed internships at the MRC Cognition and Brain Sciences Unit in Cambridge, UK and the Max Planck Institute for Psycholinguistics in Nijmegen. Upon graduation he was hired for a joint PhD position at the Institute for Logic, Language and Computation, University of Amsterdam and the Donders Institute for Brain, Cognition and Behaviour, Radboud University, the result of which you are presently reading.

Titles in the ILLC Dissertation Series:

ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language

ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax

ILLC DS-2009-03: **Brian Thomas Semmes**
A Game for the Borel Functions

ILLC DS-2009-04: **Sara L. Uckelman**
Modalities in Medieval Logic

ILLC DS-2009-05: **Andreas Witzel**
Knowledge and Games: Theory and Implementation

ILLC DS-2009-06: **Chantal Bax**
Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.

ILLC DS-2009-07: **Kata Balogh**
Theme with Variations. A Context-based Analysis of Focus

ILLC DS-2009-08: **Tomohiro Hoshi**
Epistemic Dynamics and Protocol Information

ILLC DS-2009-09: **Olivia Ladinig**
Temporal expectations and their violations

ILLC DS-2009-10: **Tikitu de Jager**
"Now that you mention it, I wonder...": Awareness, Attention, Assumption

ILLC DS-2009-11: **Michael Franke**
Signal to Act: Game Theory in Pragmatics

ILLC DS-2009-12: **Joel Uckelman**
More Than the Sum of Its Parts: Compact Preference Representation Over Combinatorial Domains

ILLC DS-2009-13: **Stefan Bold**
Cardinals as Ultrapowers. A Canonical Measure Analysis under the Axiom of Determinacy.

ILLC DS-2010-01: **Reut Tsarfaty**
Relational-Realizational Parsing

ILLC DS-2010-02: **Jonathan Zvesper**
Playing with Information

ILLC DS-2010-03: **Cédric Dégremont**
The Temporal Mind. Observations on the logic of belief change in interactive systems

ILLC DS-2010-04: **Daisuke Ikegami**
Games in Set Theory and Logic

ILLC DS-2010-05: **Jarmo Kontinen**
Coherence and Complexity in Fragments of Dependence Logic

ILLC DS-2010-06: **Yanjing Wang**
Epistemic Modelling and Protocol Dynamics

ILLC DS-2010-07: **Marc Staudacher**
Use theories of meaning between conventions and social norms

ILLC DS-2010-08: **Amélie Gheerbrant**
Fixed-Point Logics on Trees

ILLC DS-2010-09: **Gaëlle Fontaine**
Modal Fixpoint Logic: Some Model Theoretic Questions

ILLC DS-2010-10: **Jacob Vosmaer**
Logic, Algebra and Topology. Investigations into canonical extensions, duality theory and point-free topology.

ILLC DS-2010-11: **Nina Gierasimczuk**
Knowing One's Limits. Logical Analysis of Inductive Inference

ILLC DS-2010-12: **Martin Mose Bentzen**
Stit, Iit, and Deontic Logic for Action Types

ILLC DS-2011-01: **Wouter M. Koolen**
Combining Strategies Efficiently: High-Quality Decisions from Conflicting Advice

ILLC DS-2011-02: **Fernando Raymundo Velazquez-Quesada**
Small steps in dynamics of information

ILLC DS-2011-03: **Marijn Koolen**
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

ILLC DS-2011-04: **Junte Zhang**
System Evaluation of Archival Description and Access

ILLC DS-2011-05: **Lauri Keskinen**
Characterizing All Models in Infinite Cardinalities

ILLC DS-2011-06: **Rianne Kaptein**
Effective Focused Retrieval by Exploiting Query Context and Document Structure

ILLC DS-2011-07: **Jop Briët**
Grothendieck Inequalities, Nonlocal Games and Optimization

ILLC DS-2011-08: **Stefan Minica**
Dynamic Logic of Questions

ILLC DS-2011-09: **Raul Andres Leal**
Modalities Through the Looking Glass: A study on coalgebraic modal logic and their applications

ILLC DS-2011-10: **Lena Kurzen**
Complexity in Interaction

ILLC DS-2011-11: **Gideon Borensztajn**
The neural basis of structure in language

ILLC DS-2012-01: **Federico Sangati**
Decomposing and Regenerating Syntactic Trees

ILLC DS-2012-02: **Markos Mylonakis**
Learning the Latent Structure of Translation

ILLC DS-2012-03: **Edgar José Andrade Lotero**
Models of Language: Towards a practice-based account of information in natural language

ILLC DS-2012-04: **Yurii Khomskii**
Regularity Properties and Definability in the Real Number Continuum: idealized forcing, polarized partitions, Hausdorff gaps and mad families in the projective hierarchy.

ILLC DS-2012-05: **David García Soriano**
Query-Efficient Computation in Property Testing and Learning Theory

ILLC DS-2012-06: **Dimitris Gakis**
Contextual Metaphilosophy - The Case of Wittgenstein

ILLC DS-2012-07: **Pietro Galliani**
The Dynamics of Imperfect Information

ILLC DS-2012-08: **Umberto Grandi**
Binary Aggregation with Integrity Constraints

ILLC DS-2012-09: **Wesley Halcrow Holliday**
Knowing What Follows: Epistemic Closure and Epistemic Logic

ILLC DS-2012-10: **Jeremy Meyers**
Locations, Bodies, and Sets: A model theoretic investigation into nominalistic mereologies

ILLC DS-2012-11: **Floor Sietsma**
Logics of Communication and Knowledge

ILLC DS-2012-12: **Joris Dormans**
Engineering emergence: applied theory for game design

ILLC DS-2013-01: **Simon Pauw**
Size Matters: Grounding Quantifiers in Spatial Perception

ILLC DS-2013-02: **Virginie Fiutek**
Playing with Knowledge and Belief

ILLC DS-2013-03: **Giannicola Scarpa**
Quantum entanglement in non-local games, graph parameters and zero-error information theory

ILLC DS-2014-01: **Machiel Keestra**
Sculpting the Space of Actions. Explaining Human Action by Integrating Intentions and Mechanisms

ILLC DS-2014-02: **Thomas Icard**
The Algorithmic Mind: A Study of Inference in Action

ILLC DS-2014-03: **Harald A. Bastiaanse**
Very, Many, Small, Penguins

ILLC DS-2014-04: **Ben Rodenhäuser**
A Matter of Trust: Dynamic Attitudes in Epistemic Logic

ILLC DS-2015-01: **María Inés Crespo**
Affecting Meaning. Subjectivity and evaluativity in gradable adjectives.

ILLC DS-2015-02: **Mathias Winther Madsen**
The Kid, the Clerk, and the Gambler - Critical Studies in Statistics and Cognitive Science

ILLC DS-2015-03: **Shengyang Zhong**
Orthogonality and Quantum Geometry: Towards a Relational Reconstruction of Quantum Theory

ILLC DS-2015-04: **Sumit Sourabh**
Correspondence and Canonicity in Non-Classical Logic

ILLC DS-2015-05: **Facundo Carreiro**
Fragments of Fixpoint Logics: Automata and Expressiveness

ILLC DS-2016-01: **Ivano A. Ciardelli**
Questions in Logic

ILLC DS-2016-02: **Zoé Christoff**
Dynamic Logics of Networks: Information Flow and the Spread of Opinion

ILLC DS-2016-03: **Fleur Leonie Bouwer**
What do we need to hear a beat? The influence of attention, musical abilities, and accents on the perception of metrical rhythm

ILLC DS-2016-04: **Johannes Marti**
Interpreting Linguistic Behavior with Possible World Models

ILLC DS-2016-05: **Phong Lê**
Learning Vector Representations for Sentences - The Recursive Deep Learning Approach

ILLC DS-2016-06: **Gideon Maillette de Buy Wenniger**
Aligning the Foundations of Hierarchical Statistical Machine Translation

ILLC DS-2016-07: **Andreas van Cranenburgh**
Rich Statistical Parsing and Literary Language

ILLC DS-2016-08: **Florian Speelman**
Position-based Quantum Cryptography and Catalytic Computation

ILLC DS-2016-09: **Teresa Piovesan**
Quantum entanglement: insights via graph parameters and conic optimization

ILLC DS-2016-10: **Paula Henk**
Nonstandard Provability for Peano Arithmetic. A Modal Perspective

ILLC DS-2017-01: **Paolo Galeazzi**
Play Without Regret

ILLC DS-2017-02: **Riccardo Pinosio**
The Logic of Kant's Temporal Continuum

ILLC DS-2017-03: **Matthijs Westera**
Exhaustivity and intonation: a unified theory

ILLC DS-2017-04: **Giovanni Cinà**
Categories for the working modal logician

ILLC DS-2017-05: **Shane Noah Steinert-Threlkeld**
Communication and Computation: New Questions About Compositionality

ILLC DS-2017-06: **Peter Hawke**
The Problem of Epistemic Relevance

ILLC DS-2017-07: **Aybüke Özgün**
Evidence in Epistemic Logic: A Topological Perspective

ILLC DS-2017-08: **Raquel Garrido Alhama**
Computational Modelling of Artificial Language Learning: Retention, Recognition & Recurrence

ILLC DS-2017-09: **Miloš Stanojević**
Permutation Forests for Modeling Word Order in Machine Translation

ILLC DS-2018-01: **Berit Janssen**
Retained or Lost in Transmission? Analyzing and Predicting Stability in Dutch Folk Songs

ILLC DS-2018-02: **Hugo Huurdeman**
Supporting the Complex Dynamics of the Information Seeking Process

ILLC DS-2018-03: **Corina Koolen**
Reading beyond the female: The relationship between perception of author gender and literary quality

ILLC DS-2018-04: **Jelle Bruineberg**
Anticipating Affordances: Intentionality in self-organizing brain-body-environment systems

ILLC DS-2018-05: **Joachim Daiber**
Typologically Robust Statistical Machine Translation: Understanding and Exploiting Differences and Similarities Between Languages in Machine Translation

ILLC DS-2018-06: **Thomas Brochhagen**
Signaling under Uncertainty

ILLC DS-2018-07: **Julian Schlöder**
Assertion and Rejection

ILLC DS-2018-08: **Srinivasan Arunachalam**
Quantum Algorithms and Learning Theory

ILLC DS-2018-09: **Hugo de Holanda Cunha Nobrega**
Games for functions: Baire classes, Weihrauch degrees, transfinite computations, and ranks

ILLC DS-2018-10: **Chenwei Shi**
Reason to Believe

ILLC DS-2018-11: **Malvin Gattinger**
New Directions in Model Checking Dynamic Epistemic Logic

ILLC DS-2018-12: **Julia Ilin**
Filtration Revisited: Lattices of Stable Non-Classical Logics

ILLC DS-2018-13: **Jeroen Zuiddam**
Algebraic complexity, asymptotic spectra and entanglement polytopes

ILLC DS-2019-01: **Carlos Vaquero**
What Makes A Performer Unique? Idiosyncrasies and commonalities in expressive music performance

ILLC DS-2019-02: **Jort Bergfeld**
Quantum logics for expressing and proving the correctness of quantum programs

ILLC DS-2019-03: **András Gilyén**
Quantum Singular Value Transformation & Its Algorithmic Applications

ILLC DS-2019-04: **Lorenzo Galeotti**
The theory of the generalised real numbers and other topics in logic

ILLC DS-2019-05: **Nadine Theiler**
Taking a unified perspective: Resolutions and highlighting in the semantics of attitudes and particles

ILLC DS-2019-06: **Peter T.S. van der Gulik**
Considerations in Evolutionary Biochemistry

ILLC DS-2019-07: **Frederik Möllerström Lauridsen**
Cuts and Completions: Algebraic aspects of structural proof theory

ILLC DS-2020-01: **Mostafa Dehghani**
Learning with Imperfect Supervision for Language Understanding

ILLC DS-2020-02: **Koen Groenland**
Quantum protocols for few-qubit devices

ILLC DS-2020-03: **Jouke Witteveen**
Parameterized Analysis of Complexity

ILLC DS-2020-04: **Joran van Apeldoorn**
A Quantum View on Convex Optimization

ILLC DS-2020-05: **Tom Bannink**
Quantum and stochastic processes

ILLC DS-2020-06: **Dieuwke Hupkes**
Hierarchy and interpretability in neural models of language processing

ILLC DS-2020-07: **Ana Lucia Vargas Sandoval**
On the Path to the Truth: Logical & Computational Aspects of Learning

ILLC DS-2020-08: **Philip Schulz**
Latent Variable Models for Machine Translation and How to Learn Them

ILLC DS-2020-09: **Jasmijn Bastings**
A Tale of Two Sequences: Interpretable and Linguistically-Informed Deep Learning for Natural Language Processing