MAKING MEANING OUT OF EPISTEMICITY MARKERS (AND SCALARS):

A PSYCHOLINGUISTIC INVESTIGATION OF

INCREMENTAL PRAGMATIC INTERPRETATION

Dissertation

zur Erlangung des Doktorgrades

des Fachbereichs Humanwissenschaften

der Universität Osnabrück

vorgeleget

von

Vinicius Macuch Silva

aus

Curitiba

Osnabrück, 2022

To the Faculty of Osnabrück University:

The members of the Committee appointed to examine the dissertation of VINICIUS MACUCH SILVA find it satisfactory and recommend that it be accepted.

_____

Nicole Gotzner, Ph.D., Chair

_____

Michael Franke, Ph.D.

_____

Manfred Krifka, Ph.D.

ACKNOWLEDGMENT

Dissertations usually start with a lengthy and detailed acknowledgment of all those who made possible that the dissertation came into being in the first place. This dissertation is slightly different in that regard. I find it really hard to try and list all of those who supported me, given that there were so many of them, at so many different stages of my journey as a PhD. So what I will do is to simply try to reconstruct some of the personal processes behind this piece of academic writing, a milestone which wouldn't be nearly as important if it wasn't for all the individual interactions and meetings which have led to it.

Suffices to say that it took me some five-ish years to arrive at this manuscript – in between staring the PhD project in the summer of 2017 and moving countries and starting a new job while still writing the dissertation in the spring of 2022. This dissertation has seen affiliations with four institutions – University of Tübingen, Osnabrück University, ZAS Berlin, and University of Birmingham – not to mention a research stay at a fifth one – University of Edinburgh – summing up to a life period spent across five different cities, in two different countries. The dissertation has also seen both linguistics and cognitive science departments, where I met and interacted with researchers with wildly different conceptions of language and what a theory of language and cognition should look like. The dissertation has seen mostly joys but also, probably, a few sorrows. It has seen a sense of unacomplishment and insufficiency – one which has greatly extended its welcome in the writing of the manuscript. But it has also seen great personal and professional growth, as well as the growth of various circles of people. I think it makes sense to speak of circles of people, as many different people – from various walks of life – have helped pave the way to this particular academic milestone. On the more academic side, there have been colleagues, administrators, supervisors and

MAKING MEANING OUT OF EPISTEMICITY MARKERS (AND SCALARS):

A PSYCHOLINGUISTIC INVESTIGATION OF

INCREMENTAL PRAGMATIC INTERPRETATION

Abstract

by Vinicius Macuch Silva, Ph.D.
Osnabrück University
November 2022

In this dissertation, I investigate a central question in the modern inquiry into human language, namely that of how language users interpret utterances as they process them in real time, and what some of the pragmatic principles behind that might be. I juxtapose seemingly disparate themes with the goal of illustrating how pragmatic phenomena – which are incredibly rich and diverse by nature – can be brought together under an account of interpretation as an abductive process of evidence accumulation. My argumentation builds on four empirical studies focused on the interpretation of scalar expressions (Chapter 2), questions (Chapter 3), discourse particles (Chapter 4), and adjectives (Chapter 5). Taken as a unified collection of explorations on pragmatic processing, these studies contribute to the larger body of work in pragmatics and the language sciences in two significant ways: on the one hand, they provide new empirical insights to long discussed issues in the literature; on the other hand, they extend the theoretical and empirical coverage of experimental research on pragmatics to phenomena which have received little to no attention in the psycholinguistic literature. All in all, I argue that our understanding of linguistic pragmatic phenomena and their cognitive underpinnings can be enriched by construing variability in pragmatic interpretation as a function not only of the degree of conventionalization of the different meanings

associated with a linguistic form but also of the potential indexical relation between a form and its meaning (Chapter 6). In doing so, I highlight a still pervasive disconnect in the language sciences, namely that between propositional and non-propositional aspects of language and its use. By exploring from a cognitive standpoint themes which expand the purview of psychologically-oriented research on pragmatics to phenomena more closely related to the intersubjective nature of language use, this work underlines the interconnectedness between form and function in language. Ultimately, it also underlines the importance of looking, even if slightly, beyond the usual dichotomy between denotation and non-referentiality.

Contents

List of Tables

List of Figures

xiv

Dedication

This dissertation is dedicated to all the people who contributed,

in whatever shape or form,

to me being able to submit a dissertation

Chapter One

Incremental pragmatic interpretation: Generating and (dis)confirming expectations based on linguistic cues

1.1 Introduction

Given that my goal with this dissertation is to contribute, first and foremost, to the scientific understanding of human language use and its underlying pragmatic and cognitive principles, I start by presenting the goals of pragmatics as a field of research, as seen by one of its founding figures. In delineating the scope of the then emerging field of pragmatics in his foundational 1983 textbook, Stephen Levinson stated that:

> we can compute out of sequences of utterances, taken together with background assumptions about language usage, highly detailed inferences about the nature of the assumptions participants are making, and the purposes for which utterances are being used. In order to participate in ordinary language usage, one must be able to make such calculations, both in production and interpretation. This ability is independent of idiosyncratic beliefs, feelings and usages (although it may refer to those shared by participants), and is based for the most part on quite regular and relatively abstract

principles. Pragmatics can be taken to be the description of this ability, as it operates both for particular languages and language in general. (p. 53)

Nearly forty years afterwards, Levinson's words still faithfully represent the goals and purview of a now much more mature science of language use and linguistic communicative inference-making. Indeed, despite considerable methodological and empirical advances in the last decades, doing pragmatics is still synonymous with trying to get to the core of the "regular and relatively abstract principles" guiding human language use, which in itself is a rather monumental enterprise. Perhaps precisely because of the sheer scope of the enterprise, many different approaches and theoretical traditions have emerged in linguistic pragmatics since Levinson first characterized the field in the early 1980s. Commenting on the nature of pragmatic theories right before the turn of the 21st century, Gumperz and Levinson (1996) noted that "in different varieties of pragmatic theory, from Relevance Theory to more conservative Gricean theories, current work is addressed to explaining how almost vacuous or semantically general expressions can have determinate interpretations in particular contexts" (p. 8).

The theoretical landscape of the field has arguably not changed much since then, insofar as pragmaticians are still mostly preoccupied with how the production and interpretation of linguistic expressions can be explained in terms of an interplay between two types of factors, roughly defined: linguistic factors on the one hand – usually understood in terms of sententially compositional lexical semantic units of linguistic meaning – and contextual factors on the other – encompassing all sorts of non-linguistic information which feed into the inferential processes of language use. Pragmatic theorizing has been, therefore, very much centered on explaining how language users close the gap between the propositional meaning of linguistic expressions and the meaning of utterances as purposeful bits of linguistic behavior produced in context (or at least modeled theoretically as such).

In this dissertation, I address themes which relate to the broad scope of pragmatics just sketched, focusing on case studies from different families of pragmatic phenomena. I

frame my investigations against the backdrop of one specific scholarly tradition which has emerged as particularly influential on both empirical and theoretical work in pragmatics, namely the tradition following the work of philosopher of language Paul Grice. Indeed, Gricean reasoning, as defined below, remains to this day a popular framework for linguistic pragmatic analysis, influencing most work in the field either directly or indirectly. In the present dissertation I address both phenomena which have been at the center of pragmatic debate ever since the early philosophical investigations by Grice, as well as phenomena which are, arguably, more marginal relative to the core issues usually investigated in mainstream pragmatics. Perhaps precisely due to being at the margins of mainstream pragmatic debates, these phenomena provide particularly rich insights when framed and discussed using standard Gricean vocabulary.

In order to explore the different topics which I investigate as windows into human pragmatic reasoning, I draw on the tools of psycholinguistics and cognitive science. Importantly, this means that the main (empirical) backdrop to this dissertation is the experimental literature concerned with the psychological and cognitive basis of linguistic pragmatic behavior, though, whenever possible, I contrast such psychologically-centered work with accounts from neighboring literatures, both theoretical and empirical. Despite the growth of what is known as experimental pragmatics (see Noveck, 2018), two major issues still impede substantial advancements in psychologically-oriented pragmatic research. Indeed, as noted elsewhere in the literature, research in pragmatics suffers not only from a lack of explicit coupling between empirical data and abstract theory (Franke, 2016) but also from a lack of interconnection between different empirical and theoretical traditions (de Saussure and Schulz, 2007; Geurts, 2016). While these problems are widespread in the language and cognitive sciences more generally (Goldrick, 2022), they seem particularly harmful for a historically fragmented field like pragmatics. As matter of fact, rather unlike the encompassing characterization of pragmatics found in the words of Stephen Levinson in the early 1980s, the bulk of current psychologically-oriented work in pragmatics tends to be circumscribed to a handful of topics

derived mostly from the philosophical and logical traditions which have infused much of 20th century linguistics (de Saussure and Schulz, 2007; Gibbs and Colston, 2020), other topics receiving little if any attention in mainstream pragmatics.

In the present chapter, I start by briefly characterizing the core notions of the Gricean pragmatic framework which are of relevance to the work presented in this dissertation, discussing not only how they are operationalized in the different empirical chapters of the dissertation but also what some of their inherent limitations as explanatory accounts might be. I then introduce and briefly discuss the remaining theoretical notions which serve as the basis for the bulk of the empirical work reported in this dissertation. In the process of doing so, I establish the motivation as well as the main conclusions of this thesis.

1.2   Gricean pragmatics and its core principles

Most scientific incursions into linguistic pragmatic themes usually start with at least some loose reference to Paul Grice, his maxims of conversation, and his cooperative principle. Indeed, such is the foundational influence of Gricean thought on topics of pragmatic interest that his philosophical notions of conversational language use, first put forward in the 1960s, continue to motivate work across various scientific disciplines concerned with pragmatic reasoning and language use. Grice's notions are to some extent almost mythological within pragmatics, despite their striking simplicity – but also, as has been argued by many commentators, their manifest generality (see, e.g., Dänzer (2020) and Ilie (2002)).

Perhaps Grice's most relevant contribution to modern pragmatics is the concept of *implicature*, which he defined as the *implicated* meaning one arrives at on the basis of an utterance which is different from what is (literally) said or uttered. According to Grice, implicatures are either *conventional*, in the sense that they can be derived on the basis of the conventional meaning associated with the linguistic elements in an utterance, or *conversational*, in the sense that they are "essentially connected with certain general features of discourse" (Grice,

1975, p. 45). He illustrates the difference between these two forms of implicated meaning with examples. In "He is an Englishman; he is, therefore, brave", Grice observes that the utterer commits themselves to it being the case that the person they are talking about being brave is a consequence of them being an Englishman. Crucially, the derivation of such an implicated meaning follows from the meaning of the words in the utterance, in particular the meaning of *therefore*. Grice contrasts the above utterance to the reply someone gives when asked by a friend about how another friend is getting on their job: "Oh quite well, I think; he likes his colleagues, and he hasn't been to prison yet". In this case, whatever the implicated meaning of "and he hasn't been to prison yet" may be, it is clear that the utterer means something distinct from what they say, perhaps that they are surprised the friend is doing so well, or that they think the friend could be doing much worse.

Grice claims that a general principle underlies the derivation of conversational implicatures as in the latter example above, what he has termed the cooperative principle. In essence, the cooperative principle states that participants in a conversational exchange are expected to "make [their] conversational contribution such as is required, at the stage at which it occurs, by the accepted purposes or direction of the talk exchange in which they are engaged" (Grice, 1975, p. 45). What then follows from this general principle is the expectation that interlocutors adhere to what have come to be known as (Gricean) conversational maxims, a set of heuristics or norms assumed to constrain language use alongside other non-conversational norms.

The maxims originally put forward by Grice himself relate to the quantity, quality, and relation of what is said in an utterance as well as to the manner with which something is said, all of which, according to him, "will, in general, yield results in accordance with the Cooperative Principle" (Grice, 1975, p. 45). While Grice's own wording of the maxims is infamous for being vague and general, and while the exact theoretical status of his (sub-)maxims varies with different reinterpretations of his original principles, there is at least some degree of consensus in the literature that the core assumptions underlying the cate-

gories introduced by him are relevant for the characterization of the pragmatic processes involved in the generation of implicatures (see, *inter alia*, Ariel, 2016). Disagreement stems most notably from the so-called Post-Gricean tradition, where the process of utterance interpretation is framed in altogether different terms (see, e.g., Sperber and Wilson, 1986; Wilson, 2019). While the empirical investigations reported in this thesis may be framed in both (Neo-)Gricean and Post-Gricean terms, I formulate my discussions using primarily Gricean vocabulary, in line with the interpretation put forward by Dänzer (2020) whereby "the goal of Gricean pragmatics is to offer rationalizing explanations of aspects of utterance interpretation, and the role of Gricean derivations is to provide the central elements of these explanations" (p. 704).

Crucially, according to this view, rationalizing explanations such as the ones offered by Gricean pragmatic accounts constitute psychological explanations to the extent that they construe interpretation "as the manifestation of a rational disposition, that is, a disposition to do or believe what is rational in light of one's mental states" (Dänzer, 2020, p. 691). Whether interpretation itself, as a process, is best construed in terms of intention recognition or some form of commitment or social action ascription is a foundational question that remains open and is not directly addressed by the studies presented in this dissertation. In fact, one can view pragmatic interpretation in the terms discussed above while also assuming that inferencing is primarily a normative process, contrary to many interpretations of the Gricean program. Geurts (2019) notes that some degree of "mentalism can be granted while maintaining that we get much of our pragmatic business done without attributing mental states to each other" (p.7). This seems to be compatible with Dänzer's (2020) interpretation of the Gricean program, if not to say with Grice's own words in his documented writings.

Given that in this dissertation I'm interested in empirically assessing the suitability of rationalizing explanations which have been proposed to account for the phenomena discussed in chapters 2 to 5, I'll start by introducing these proposed explanations. Two of them can be cast in terms of the Gricean categories of quantity and manner. As the name suggests, the

category of quantity relates to the quantity of information which is provided in an utterance. Grice originally put forward two quantity maxims, stating that interlocutors are expected to "Make [their] contribution as informative as is required (for the current purposes of the exchange)" and to "Do not make [their] contribution more informative than is required". While general, this notion of quantity can be somewhat more concretely instantiated in different assumptions, such as the assumption that the informativity of a given linguistic expression is dependent on their semantic strength relative to other relevant linguistic alternatives. In Chapter 2, I present a study couched precisely in the assumption that the informativity of a scalar expression in context might be pragmatically relevant for the interpretation of an utterance containing such an expression.

As for the category of manner, unlike quantity, it relates not to what is said but rather to how what is said is said. Grice captured the core intuition behind the notion of manner in his supermaxim "Be perspicuous", which he noted could be instantiated in various maxims such as "Avoid obscurity of expression", "Avoid ambiguity", "Be brief (avoid unnecessary prolixity)", and "Be orderly". In Chapter 3, I present a study couched in this general notion of manner, more specifically in the assumption that alternative formulations of an utterance might be pragmatically relevant for interpretation in light of background assumptions which might bias an interpreter to expect one form over the other.

While the core theoretical assumptions entertained in chapters 2 and 3 can be traced back to the Gricean notions of quantity and manner, they are instantiated in more concrete assumptions related to the specific phenomena investigated in each of those chapters. In the other two empirical chapters of this dissertation, chapters 4 and 5, I present studies which can be couched in the notion of conventional implicatures, which, as discussed earlier, are said to operate differently from their conversational counterparts. The concept of conventional implicature has been further developed, most notably, by Potts (2007), who notes that "CIs define a dimension of meaning that, though conventionally encoded, is separate from the primary semantic content" (p. 666).

Potts notes, further, that "the degree of separation varies from analysis to analysis, but this *multidimensionality* is in evidence throughout" (p. 666). According to his account, despite being separate from the primary semantic content of an expression, conventional meanings are not pragmatic, in the sense that they do not function like conversationally implicated meanings, whose defining feature is that they are calculable in context on the basis of general principles of conversation. Feng (2011, 2010) provides an alternative account where conventional meanings are assumed to be pragmatic, or at the very least both semantic and pragmatic, in the sense that a conventional implicature:

> "is so named because it involves both linguistic and contextual information. It is conventional because it is associated with the conventional linguistic meaning of a certain expression, from which one would know something about what the speaker would mean given the knowledge of the language he speaks. It is implicated rather than said because its full content requires contextual information, and does not affect the truth conditions of the utterance." (Feng, 2010, p. 110)

In this dissertation, I will assume not only the core distinction between conversational and conventional implicatures commonly assumed in the literature but also that conventionally implicated meanings are pragmatic, akin to their treatment in Feng's account. This begs distinguishing pragmatic meanings from semantic ones, a task which has been at the core of pragmatics since its very inception. While not under any account an easy task, as evidenced by the fact that is has troubled language scientists for decades, for the purposes of this dissertation one can define as 'semantic' those components of meaning which are truth-conditional, that is, those aspects of meaning which can be captured in terms of truth conditions. While this is a relatively uncontroversial definition of semantic meaning, at least for the purposes of theoretical abstraction, defining pragmatic meaning quickly turns to a hair-splitting endeavor, one laden not only with theoretical friction but also partisan sentiments. Many terms with diverging definitions and interpretations have been used in the

literature to encompass whatever pragmatic meanings are supposed to be vis-à-vis semantic meanings (see Sander (2022) for a discussion) – secondary meanings or content (e.g. Gutzmann and Turgay, 2019), use-conditional meanings (e.g. Gutzmann and Gärtner, 2013), non-at-issue meanings (e.g. Tonhauser et al., 2013), procedural meanings (e.g. Blakemore, 1987), among others. For my present purposes, I define as 'pragmatic' those components of meaning which are non-truth-conditional, that is, those aspects of meaning which cannot be straightforwardly captured in terms of truth conditions. While a negative definition is hardly ever a useful one, the definition proposed here is meant to illustrate the fuzziness inherent to pragmatic phenomena, which are perhaps collectively best characterized in terms of how they function as probabilistic procedures over otherwise truth-conditional meanings. From these definitions it follows that the boundaries between semantic and pragmatic meanings are fuzzy, perhaps much more of a product of theoretical abstraction than an actual empirical state of affairs. Still, for the sake of simplicity, suffices to say that pragmatic, non-truth-conditional meanings can be thought to lie on a gradient with semantic, truth-conditional meanings.

In other words, while I assume semantic meanings to be primarily propositional in nature, I assume pragmatic meanings to be non-propositional, whatever their exact meaning contribution may be. Crucially, both semantic and pragmatic meanings can have propositional implications for interpretation, and both semantic and pragmatic meanings can be conventionally associated with a given linguistic form, at least according to the definitions above. This entails that the propositional implications of a given meaning, whether the meaning is semantic or pragmatic, might depend on its degree of conventionalization with a particular linguistic form. At the same time, the propositional implications of a given meaning might depend not only on whether it is semantic or pragmatic but also on whether it leads to any actual pragmatic computations in context, as would be canonically the case with conversational implicatures and as is the case with conventional implicatures according to an account like Feng's (2011; 2010).

All in all, this paints a picture of interpretation where the nature of any given inference – whether an implicature in the Gricean sense or something else – is dependent not so much on the type of meaning a particular linguistic form encodes but rather on the propositional and non-propositional implications that meaning has for the task of interpretation. This picture captures both the fact that a single form can carry multiple, potentially quite distinct meanings as well as the fact that each of those meanings can have a stronger or weaker impact on interpretation in context. Ultimately, then, whatever the nature of the relationship between different meaning components of a given linguistic form, it is important to distinguish the potentially pragmatic, i.e., non-truth-conditional, nature of a linguistically encoded meaning from any pragmatic, i.e., inferential, derivation processes potentially at play during interpretation. While the former sense of 'pragmatic' encapsulates a semiotic relation between a form and a meaning, one which is expected to have implications for interpretation, the latter encapsulates a reasoning process, one which is amenable to rationalizing explanations and which can be more or less concretely instantiated in different cognitive processes.

The relationship between the nature of a form-meaning mapping and its impact on interpretation is discussed in detail in Chapter 6, the general discussion of this thesis. My main conclusion in that chapter is that meaning conventionalization seems to be a key factor impacting the likelihood of a linguistic cue giving rise to an inference in context, regardless of the exact meaning that cue contributes. This is said to be the case particularly for forms which encode non-truth-conditional meanings, which might be, a priori, less likely to have concrete propositional implications for interpretation, unlike forms which encode truth-conditional meanings, which are, by definition, relevant for parsing. In fact, under the assumption that meaning conventionalization is graded, the less conventionalized a given pragmatic meaning is, the less likely it is to trigger a pragmatic derivation in the first place. I discuss this conclusion with regard to the phenomena investigated in three of the four empirical chapters of this thesis, all of which deal with the interpretation of utterances containing so-called epistemicity markers. Before moving on to define the concepts of prediction and

expectation, which lay the psycholinguistic foundation of this thesis, I discuss why markers of epistemicity with conventionalized pragmatic meanings provide particularly interesting case studies for the study of pragmatic interpretation, which motivates their study in this thesis.

1.3   Epistemicity, indexicality, and inferencing

Topics which are classically considered to fall under the auspice of pragmatics are, among others, implicature, presupposition, reference, and deixis. While in the domain of deixis the context of utterance is assumed to be logically necessary for the interpretation of deictic expressions such as *I* or *tomorrow*, in the case of most other pragmatic phenomena the meaning of a pragmatically relevant linguistic expression is usually modeled as being constrained but not in itself determined by the context of utterance. For instance, in the case of so-called scalar implicatures, which are discussed in detail in Chapter 2, the meaning of a scalar expression like the quantifier *some* can be constrained via an inferential process whereby an upper-bounded interpretation (i.e. some but likely not all) is derived from an otherwise lower-bounded interpretation (i.e. some and potentially all), given some form of contextual support for the former. However interesting the inferential process behind the derivation of such an implicated meaning might be, what's at stake is the fact that the meaning which is enriched is a referential one, such that it refers to some entity, event or more generally a state of affairs – in the present case the size of a particular set of interest – which can be rather straightforwardly captured using truth conditions. In other words, according to the definitions of semantic and pragmatic meaning discussed above, the meaning of *some* can be regarded as semantic, i.e., as truth-conditional, given that it can be captured in terms of whether or not it satisfies certain truth conditions. What is pragmatically relevant about the interpretation of a scalar expression like *some* is the potential enrichment or constraining of its meaning from a lower-bounded one (i.e. some and potentially all) to an upper-bounded

one (i.e. some but likely not all), which is assumed to consist of a pragmatic derivation process, one which in a Gricean framework is construed as being related to considerations of quantity or informativity, as discussed in section 1.2 of the current chapter.

Despite involving potentially different underlying pragmatic processes, other pragmatic phenomena like presuppositions and reference are no different than implicatures when it comes to the nature of the meanings the pragmatic processes are said to operate over: the former are usually concerned with some state of affairs which may or may not be in the common ground between two interlocutors (e.g. whether something has *stopped* happening or whether it is happening *again*), the latter being usually concerned with concrete referents, whether entities or abstract states of affairs (e.g. a *he* or *it* which is mapped onto a real world entity like a person or a situation). Said differently, classic pragmatic phenomena tend to be concerned with the mapping between linguistic expressions and some referential state of affairs, such as the mapping between a pronoun in a referring expression and its actual referent (e.g. 'he' in "What was *he* talking about?"), or the mapping between a verb and a presupposed state of affairs it may refer to (e.g. 'stop' in "They *stopped* eating sugar altogether"). Even in otherwise purely referential scenarios, however, language use tends to involve more than the mere linkage between linguistic signs and entities in a physical or imagined world. It more often than not involves the expression of evaluations and expectations with regard to what is being referred. Linguistically, this is achieved via a multitude of strategies, including the usage of structures and devices which are not classically thought to fall under the purview of pragmatics. Modal devices are among some of these strategies.

Indeed, modality is a central theme in linguistic research, one which cuts across many subfields and approaches and which has received as much scholarly attention as classic topics in grammar. Two of its sub-types, epistemic modality and evidentiality, are intrinsically related to the evaluation – both subjective and objective – of the truth, plausibility or even relevance of states of affairs. Interestingly, these two sub-types of linguistic modality are also intrinsically related to one another (Cornillie, 2009; Ricci and Rossari, 2018), such

that markers of modality commonly express meanings which blur the line between what is prototypically considered to be the marking of knowledge – the domain of epistemic modality, sometimes referred to as 'epistemicity' – and the marking of information source – the domain of evidentiality (Gray and Biber, 2012). Given this picture, referring to the linguistic marking of knowledge and the linguistic marking of information source in an encompassing manner is oftentimes relevant, yet there seems to be no common term in the literature encompassing both these flavors of linguistic modality, with the exception of the term 'commitment', which, as discussed in de Brabanter and Dendale (2008), is used in a variety of different ways in closely related fields. In this thesis I, therefore, use the term 'epistemicity' to cover this usage and to avoid any strong terminological implications.

Relevant for my purposes in this dissertation is the fact that, by definition, linguistic markers of epistemicity can be construed as overt cues to information states, both subjective and objective, which means that, from a theoretical perspective, they can be expected to serve as pragmatic cues to semantic information and thus, by extension, as potentially strong constraints on interpretation. In fact, many of these markers have undergone processes of pragmaticalization whereby their primary function has become to signal what the relationship between an utterance and its context might be (Diewald, 2011; Traugott, 2012, 1995), as opposed to contributing any truth-conditional meaning to an utterance. Regardless of the extent to which they may be fully pragmaticalized, epistemicity devices seem particularly relevant for pragmatic analyses exactly because they *index* how information in an utterance might interact with information which is (assumed to be) part of the common ground. That is to say that, unlike in the case of linguistic signs which can be used to directly establish reference to some state of affairs in the world, including deictic expressions, indexical pragmatic devices simply 'point' to semantic information which itself carries referential meaning in context, potentially modulating its import in the context of utterance by indicating either how the utterer evaluates such information or how it might relate to other information not introduced in the utterance itself. In short, given that epistemicity devices are assumed

to facilitate interpretation by means of indexing how (part of an) utterance might interact with its context, they seem like ideal testbeds for theories of how pragmatic interpretation is derived in context.

Generally speaking, the findings reported in chapters 3 to 5 show that the usage of epistemicity markers can indeed modulate interpretation by virtue of what their conventionalized meanings signal in context. However, as discussed in detail in Chapter 6, the conclusion that the pragmatic impact of a given linguistic marker varies as a function of the conventionlization of its pragmatic meaning is confounded by several other factors, namely the linguistic nature of the specific device at hand (e.g. whether a lexical or a (morpho)syntactic device), the potential polysemous character of a given device, as well as the nature of any other meanings which might compete with the pragmatic meaning of a device (i.e. whether they are also pragmatic or semantic in nature).

All in all, given that epistemicity phenomena tend to involve not only cases of polysemy but also polyfunctionality, investigating such phenomena becomes an intricate but at the same time elucidating task in light of the fact that their interpretation can be accounted for in terms of potentially different interpretive processes. As discussed in the next section, the question of how meaning can be derived from conventional pragmatic devices like epistemicity markers becomes all the more interesting when attempting to account for interpretation in terms of how it unfolds incrementally. In this regard, couching accounts of pragmatic interpretation in psycholinguistic theories of predictive processing may help better understand what the nature of the underlying interpretive processes is.

1.4 Prediction and predictability in online language processing

Prediction is a fundamental notion in modern cognitive science (e.g. Bubić et al., 2010; Clark, 2013), one which has made its way into psycholinguistic research (see, *inter alia*, Kuperberg and Jaeger, 2016). The core intuition behind the notion of prediction is that humans, as cognitive agents, are able to predict upcoming information on the basis of different sorts of cues, from perceptual and social cues to linguistic ones (e.g., Kamide et al., 2003; Ridderinkhof, 2017; Winkler et al., 2009; see also Litwin and Miłkowski, 2020). In this sense, one can speak of cues being more or less predicable, where predictability can be taken as the expected likelihood of a given cue being predicted in light of previous cues. While prediction has become a key tenet of recent psycholinguistic research, its exact role in language processing is still debated, on the one hand with regard to its centrality or necessity (Huettig and Mani, 2016), and on the other hand with regard to its nature, that is, whether it involves strict anticipation or pre-activation of material or some form of expectation via backward integration (see Kutas et al. (2011) and Ferreira and Chantavarin (2018)). While evidence for prediction exists across all levels of linguistic representation, there is a growing consensus that it might not be as central to language comprehension as is often claimed in the literature (Huettig and Mani, 2016), at least not when construing prediction as strict pre-activation of linguistic material. Adding to this point and speaking to the fact that distinguishing between different forms of prediction might not be necessary when assuming a unified generative system which operates at different levels of representation, multidimensional models featuring both prediction and integration have been proposed as parsimonious accounts of incremental language processing (e.g., Ferreira and Chantavarin, 2018; Kuperberg and Jaeger, 2016). Importantly, these models can not only account for prediction but also for its absence: anticipating information at lower levels of representation is not expected to occur if it's not supported by sufficient evidence from higher levels of

representation; at the same time, information from lower levels of representation actively contributes to the build-up of evidential support at higher levels of representation, such that richer representations are more likely to give rise to (strong) predictions.

When it comes to pragmatic processing, expecting a particular word can be taken to mean that comprehenders have generated predictions about the likelihood of that word – or at least its rough meaning – serving as a pragmatically felicitous continuation to an unfolding discourse model, as per expectancy/ constraint-based approaches to pragmatic processing (Degen and Tanenhaus, 2019; Rohde and Kurumada, 2018). Regarding, more specifically, the reasons as to why comprehenders might predict linguistic material based on pragmatic expectations, aside from the theoretical assumptions linked to specific types of pragmatic inference, there is the general assumption that linguistic cues can be pragmatically relevant for interpretation as they can, at least in principle, provide enough evidential support to trigger an inference. In this regard, a general finding from the studies discussed in this thesis is that pragmatic processing seems to unfold based on constraints which are, on the one hand, more varied and context-dependent compared to what is usually assumed in the literature. On the other hand, the constraints discussed in the different chapters of this thesis all seem to share a common explanatory basis, namely the assumption that inferencing and predictive processing advance based on the accumulation of evidential support for a particular interpretation, at least in the case of phenomena which are canonically treated as (Gricean) implicatures. With these considerations in mind, in the sections below I summarize the main assumptions and results of each empirical chapter of this thesis.

1.5   Chapter 2

In Chapter 2, I present a study focused on a well-studied class of pragmatic inferencing known as scalar implicatures. From a theoretical perspective, the derivation of an inference cued by a scalar expression rests on the assumption that language users reason about linguistic

alternatives, more specifically lexical alternatives which are construed as being ordered on a scale according to their relative semantic strength[1]. Crucially, it follows from this account that the relationship between alternatives on a particular lexical scale – as between the quantifiers *some* and *all* on a scale of quantity – is pragmatically relevant for interpretation as different items on the scale are more or less informative in context depending on their semantic status relative to one another. As such, when interpreting an utterance with a semantically weak scalar, one might infer, given (Gricean) pragmatic expectations, that the expression is being used informatively and thus that it's the strongest possible scalar alternative in that given situation. For instance, when interpreting an utterance with a scalar expression like the quantifier *some* one might therefore infer that, unless suggested otherwise by other cues in the signal or in the discourse context, the set of interest is that of some but likely not all, given that the stronger alternative *all* was not actually uttered.

In the study reported in Chapter 2, I investigated the interpretation of utterances containing the German scalar quantifiers *einige* and *alle*, counterparts to English *some* and *all*. As has been shown before, pragmatic expectations about the informative usage of scalars affect incremental interpretation in real-time, such that scalar quantifiers can serve as predictive cues to meaning in online processing. Indeed, results from previous psycholinguistic studies show that comprehenders are able to predict subsequent material in a sentence upon processing a quantifier like *some* (e.g. Degen and Tanenhaus, 2016; Nieuwland et al., 2010), drawing scalar implicatures when these are licensed in context. However, previous research has not looked at the *relative* predictability of subsequent material in a sentence when processing a lower-bounded scalar – which might give rise to an implicature given its informativity in

---

[1]While it is debated what exact form of inferencing best characterizes the derivation of scalar implicatures (see e.g. Gotzner and Romoli, 2022), the current empirical picture suggests that lexical scales play an important role in the interpretation of scalar terms, therefore supporting a characterization of scalar inferences as generalized conversational implicatures dependent primarily on the conventionalized (semantic) meaning of scalars as opposed to *ad hoc* inferencing in context, which would instead support a characterization in terms of particularized implicatures/ explicatures.

context – as opposed to an upper-bounded scalar, which might serve as a predictive cue to meaning without giving rise to a pragmatic inference. In other words, despite previous research having shed light on the circumstances under which processing a scalar like the quantifier *some* might give rise to an enriched pragmatic interpretation, the relative predictability of either an enriched or a non-enriched interpretation of a sentence containing a lower-bounded scalar has not been directly investigated.

In the setup reported in Chapter 2, I investigated this question using a self-paced reading task embedded in a picture-verification task. I started by investigating whether language users have biases similar to those expected under an informativity account when producing utterances containing scalar quantifiers. Using visual stimuli depicting two arrays of geometric shapes which varied in terms of their color, shape, and size I showed that, when prompted to produce descriptions of such stimuli by picking a scalar quantifier (i.e. *some* and *all*), a shape term (i.e. *circles*, *squares*, and *triangles*), and a property term (e.g. *orange*, *blue*, *yellow*, and *green*), German speakers prefer producing descriptions with quantifiers which are semantically stronger, and thus pragmatically more informative, than their scalar alternatives, e.g., "*All* of the triangles in the picture are yellow", when all triangles in the picture were yellow. They also prefer producing descriptions which are maximally distinct from other possible descriptions of the same picture, e.g., "*Some* of the circles in the picture are *orange*", when all triangles in the picture were yellow and half of the circles were yellow while the other half were orange. Importantly, these biases are in line with those intuitively expected under an informativity account, whereby scalars are used in informative ways given their relative semantic strength.

In a picture-verification task using the same visual stimuli and the same sorts of descriptions as in the production task, I investigated whether such biases found in the production of descriptions containing scalar quantifiers are reflected in the comprehension of the descriptions, more specifically when people parse them incrementally. Crucially, the task involved reading a description of a picture one word at a time after having seen the picture, the pair-

ings between pictures and their respective descriptions being such that the quantifiers in the descriptions were expected to lead to either more or less predictable mentions of a particular shape array (e.g. triangles) when considering an incremental unfolding of the utterance, as when the description is read one word at a time. The key assumption underlying this setup was that, after looking at the target picture and reading the quantifier in the referring expression, more than one informative sentence continuation was available per visual scene in so-called unbiased scenarios – scenarios where the description was compatible with both shape arrays depicted in the picture, whereas in so-called biased scenarios – scenarios where the description was compatible with only one shape array depicted in the picture – only one informative continuation was available. Therefore, in the biased conditions, participants were expected to generate strong predictions about a specific sentence continuation, which implies strong expectations about the to-be-read shape term in the unfolding sentence. In the unbiased conditions, on the other hand, participants were expected to generate weaker predictions about one or the other possible sentence continuations, which implies weaker relative expectations about the to-be-read shape term in the unfolding sentence. Whether a given prediction A or B was confirmed, participants were expected to be surprised by what they read, given that they did not have strong expectations about a given term X or Y in the first place. This account was termed pragmatic surprisal, in line with previous work investigating predictability by surprisal in (predictive) sentence processing.

Contrary to what was expected under the pragmatic surprisal account, I found that comprehenders did not strongly predict the shape term in the referring expressions when the quantifier strongly biased a particular shape term, i.e., when the unfolding description was compatible with only one shape array in the target picture. In fact, the empirical results showed the opposite pattern to that predicted by pragmatic surprisal, whereby comprehenders read the shape term faster in unbiased scenarios compared to biased scenarios, suggesting that they generated stronger predictions about the shape term in the former rather than in the latter.

Ultimately, I concluded from these results that predictability in the incremental interpretation of utterances containing scalar quantifiers is linked not only to purely informativity-based prediction but also to other pragmatic constraints, such as, in the case of this particular study, one derived from a pressure to integrate crucial semantic information incrementally during processing. This is in line with constraint-based accounts of language comprehension which assume not only that multiple sources of information need to be integrated online during processing but perhaps more importantly that the weight of different constraints varies depending on the specific processing demands as well as on the larger discourse and communicative context (Degen and Tanenhaus, 2019). The findings also reflect the findings from other studies on online pragmatic processing which have shown that pragmatic inferencing, including scalar inferencing cued by quantifiers, shows variable time courses and strong context-dependence (Bergen and D. J. Grodner, 2012; Huang and Snedeker, 2018; Urbach et al., 2015).

1.6   Chapter 3

In Chapter 3, I present a study focused on a class of pragmatic inferencing known as manner implicatures. From a theoretical perspective, the derivation of an inference cued by a particular utterance form rests on the same assumption as that explored in Chapter 2, namely that language users reason about linguistic alternatives. In the case of manner implicatures, however, unlike with their scalar counterparts, inferencing is not grounded in reasoning about lexical alternatives and their informativeness in context. Rather, it's grounded in reasoning about alternative utterance formulations and what processing one formulation versus another might mean pragmatically, more specifically in the case at hand what the morphosyntactic form of a negated polar question might signal about the epistemic stance of the questioner regarding the content of the question. Crucially, it follows from this account that the relationship between alternative morphosyntactic forms of a negated polar question – for instance,

20

in English, the auxiliary verb construction with a contraction *Didn't you get the ticket?* as opposed to its uncontracted counterpart *Did you not get the ticket?* – is pragmatically relevant for interpretation as the questioner stance implicated by each formulation might be more or less felicitous depending on any prior expectations the interpreter might have about the information state of the questioner, meaning that the form signaling the questioner's commitment to the truth of the content communicated in the question should be more felicitous given an expectation that the questioner believes in the truth of the proposition, while the form signaling no commitment to the truth of that same state of affairs should be more felicitous given the absence of any expectation regarding the questioner's epistemic state. As such, when interpreting a negated polar question with a form associated with a neutral questioner stance – namely an uncontracted auxiliary construction as in *Did you not buy a ticket?* – in the absence of any particular expectation about the epistemic state of the questioner, one might infer, given pragmatic expectations about how the utterance is formulated and what that formulation might signal in context, that the questioner does not actually know whether the situation they are enquiring about is true or not. On the other hand, when interpreting a negated polar question with a form associated with the questioner having a positive epistemic stance on the issue – namely a contracted auxiliary construction as in *Didn't you buy a ticket?* – having an expectation that they believe in the truth of the content communicated in the question, one might infer, given the same pragmatic expectations as before, that the questioner expects the situation to be true. It is not clear, however, what exact predictions follow from this account when it comes to incremental interpretation. In fact, the relationship between the intuitions above and the actual processing expectations that might emerge in the incremental interpretation of questions with biases, as they are called in the literature, can be conceptualized in different ways. On the one hand, it might be that the processing of a felicitous question form is facilitated compared to the processing of an infelicitous question form – which is a hypothesis that assumes pragmatic infelicity to be a strong constraint on the interpretation of a question with bias. This entails prag-

matic facilitation at the question form itself, which spans several words both in the English constructions (*Did you not* buy a ticket? vs. *Didn't you* buy a ticket?) as well as in the German ones (*Hast du kein* Ticket gekauft? vs. *Hast du nicht ein* Ticket gekauft?). On the other hand, it might be that processing a felicitous question form facilitates the processing of subsequent semantic material in the sentence compared to processing an infelicitous question form – which is a hypothesis that, compared to its counterpart above, assumes pragmatic infelicity to be a weaker constraint on the interpretation of a question with bias. This entails pragmatic facilitation at a semantic element following the question form. Linearly, the first semantic cue following the auxiliary construction is the main verb in the English constructions (Did you not *buy* a ticket? vs. Didn't you *buy* a ticket?) and the noun in the German constructions (Hast du kein *Ticket* gekauft? vs. Hast du nicht ein *Ticket* gekauft?). While before collecting and analyzing any data I entertained a hypothesis compatible with a view of pragmatic felicity as a strong constraint on the interpretation of question with biases, after having collected and analyzed the data as originally planned I revised my predictions according to the weak constraint hypothesis and collected a new data set. Below I report the findings of the study with reference to the weak constraint hypothesis, i.e., one which predicts pragmatic facilitation at a semantic cue following the question form as opposed to at the question form itself.

In the study reported in Chapter 3, I investigated the incremental interpretation of questions with biases in German and in English. Even though the topic has received very little attention in psycholinguistic research, it has been shown that both English and German speakers formulate polar questions differently depending on their information state regarding the content of the question, such that, when prompted to select between different formulations in an experimental task which manipulates both epistemic and evidential information regarding the state of affairs in the question, they show clear biases for particular forms in different epistemic contexts (Domaneschi et al., 2017). In English, in situations constructed not to elicit any particular prior belief about the state of affairs targeted in the question, peo-

ple prefer an uncontracted auxiliary construction (i.e. *Did you not buy a ticket?*), whereas in situations constructed to elicit an expectation that the state of affairs in the question was true they prefer a contracted auxiliary construction (i.e. *Didn't you buy a ticket?*). In German, in situations constructed not to elicit any particular prior belief about the state of affairs in the question, people prefer a construction with a negated determiner (i.e. *Hast du kein Ticket gekauft?*), whereas in situations constructed to elicit an expectation that the state of affairs in the question was true they prefer a construction with sentential negation (i.e. *Hast du nicht ein Ticket gekauft?*). What this shows is that, in English and German, alternative morphosyntactic forms of a polar question correlate with different questioner stances, despite, at face value, the different forms having no difference in meaning. While such epistemic biases have been found to affect the production of polar questions, in Chapter 3 I investigated whether they also affect comprehension.

In the setup reported in Chapter 3, I used a self-paced reading task in which participants read negated polar questions against discourse contexts built to elicit the epistemic biases discussed above. I started by norming the discourse scenarios so as to ensure that they would elicit the relevant biases. The normed items show that, when prompted to answer a paraphrase of the polar question used in the main reading task, e.g., "Did the person in this scenario buy a ticket?", both English and German speakers show response biases which reflect the relevant questioner epistemic stances reported in the literature, namely no particular answer expectation in the neutral stance situations and an expectation that the answer is positive in the positive stance situations. Importantly, this shows that the sentences used as discourse contexts reflect the epistemic information which, together with evidential information licensing a negated question in the first place, should license the usage of the relevant question forms in each language, i.e., the contracted vs. uncontracted auxiliary construction in English and the construction with a negated determiner vs. the construction with sentential negation in German. In the reading task I then investigated whether the combination of such contextual biases with the contextually-appropriate question forms impacts incremental

interpretation. Crucially, the task involved reading a negated polar question one word at a time after reading the relevant context sentences. The key assumption underlying this setup was that, after reading the discourse context and the fronted auxiliary construction in the question, subsequent semantic material in the sentence should be more or less predictable depending on the felicity of the respective question forms in context. Therefore, in neutral stance situations, participants were expected to generate stronger predictions about the lexical cue directly following the question form when processing an uncontracted auxiliary construction in English (i.e. *Did you not buy a ticket?*) and a construction with a negated determiner in German (i.e. *Hast du kein Ticket gekauft?*), which implies stronger expectations about the to-be-read verb in the unfolding English sentence and the to-be-read noun in the unfolding German sentence. In positive stance situations, on the other hand, participants were expected to generate stronger predictions about the lexical cue directly following the question form when processing a contracted auxiliary construction in English (i.e. *Didn't you buy a ticket?*) and a construction with sentential negation in German (i.e. *Hast du nicht ein Ticket gekauft?*), which implies stronger expectations about the to-be-read verb in the unfolding English sentence and the to-be-read noun in the unfolding German sentence. This account was termed predictability by question bias, in line with the work characterizing the phenomenon as an issue of question bias.

Unlike what was expected under the account of predictability by question bias, I found that comprehenders did not strongly predict the semantic cue directly following the question form either in English or in German. However, the empirical results showed that, at least in English, comprehenders read the last semantic cue in the questions predictively depending on the question form and bias profile. In English, comprehenders read the noun faster in contracted auxiliary constructions (i.e. *Didn't you buy a ticket?*) than in uncontracted auxiliary constructions (i.e. *Did you not buy a ticket?*) when the questions appeared in positive stance contexts. In German, on the other hand, the descriptive results suggest that comprehenders read the main verb faster in constructions with a negated determiner (i.e. *Hast du*

kein Ticket gekauft?) than in constructions with sentential negation (i.e. *Hast du nicht ein Ticket gekauft?*) when the questions appeared in positive stance contexts, though this result find no quantitative support in the respective statistical model. While the facilitatory effect in the English data is in line with the general predictions of predictability by question bias, despite emerging further downstream in the sentence, the potential facilitation attested in the German data is in the opposite direction to that expected under the predictability by question bias account. All in all, what this seems to suggest is that comprehenders might be able to predict relevant semantic material downstream in a (negated) polar question when the discourse context biases a positive questioner stance, though it is unclear how exactly this bias interacts with the form of a question.

Ultimately, I concluded from these results that predictability in the incremental interpretation of negated polar questions is not strongly constrained by the felicity of the mappings between contextual (epistemic) biases and the particular (morphosyntactic) form of a question, even when postulating that pragmatic felicity constrains interpretation insofar as it facilitates the processing of semantic material after the question form itself has been processed. In fact, a stronger pragmatic constraint seems to be the overall presence of a questioner stance, given that comprehenders are able to predict semantic information in questions embedded in contexts with a positive questioner stance but not in questions embedded in contexts with a neutral questioner stance. In other words, only when the context biases a non-neutral questioner stance does the felicity of the mappings seem to impact interpretation, though in the case of German the form that facilitated processing downstream in the sentence was not the expected one. Future studies on the interpretation of questions with biases should therefore aim to further disentangle the conditions under which the incremental tracking of a bias – whether epistemic or not in nature – might be relevant for interpretation.

1.7 Chapter 4

In Chapter 4, I present a study focused on a class of pragmatic inferencing known as conventional implicatures. From a theoretical perspective, the derivation of an inference cued by the meaning of a linguistic form rests on the same core assumption explored in Chapters 2 and 3, namely that language users reason about the pragmatic implications of the meanings associated with particular linguistic forms. In the case of conventional implicatures, however, unlike with their scalar and manner counterparts, inferencing is not grounded in reasoning about alternative forms and what processing one form versus another might mean pragmatically. Rather, it's grounded in reasoning about what the meanings conventionally associated with a particular form might mean in context, more specifically in the case at hand what the non-truth-conditional meaning of a discourse particle might signal about the truth-conditional meaning of the utterance it is embedded in. Crucially, it follows from this account that the pragmatic meaning conventionally associated with a discourse particle is relevant for interpretation as it signals how the semantic meaning of its host utterance might relate to previous information in the discourse. As such, when interpreting a particle which signals a contrast between the meaning of its host utterance and that of a contextually-relevant antecedent, one might infer, given pragmatic expectations, that the content of the utterance is not congruent with the unfolding discourse, suggesting a disconfirmation or disagreement with what has been said in the relevant discourse antecedent. On the other hand, when interpreting a particle which signals an agreement between the meaning of its host utterance and that of a contextually-relevant antecedent, one might infer, given the same pragmatic expectations, that the content of the utterance is congruent with the unfolding discourse, suggesting a confirmation or agreement with what has been said in the relevant discourse antecedent.

In the study reported in Chapter 4, I investigated the incremental interpretation of utterances with discourse particles in German. The topic has received very little attention

in psycholinguistic research and previous findings are conflicting. On the one hand, there is evidence that, in Dutch, pragmatic expectations about the usage of discourse particles affect incremental interpretation in real-time (van Bergen and Bosker, 2018), such that the discourse particles *eingelijk* and *inderdaad* can serve as predictive cues to meaning in online processing. Indeed, while processing a contrast-marking particle like *eingelijk* induces comprehenders to expect a disagreement between its host utterance and a contextually-relevant antecedent, allowing them to predict a contextually-relevant referential alternative, processing an agreement-marking particle like *inderdaad* induces comprehenders to expect an agreement between its host utterance and a contextually-relevant antecedent, allowing them to predict a referent biased by the context. More recent results, however, provide no evidence that comprehenders draw inferences about linguistic material downstream in Dutch sentences with the particles *eingelijk* and *inderdaad* (Rasenberg et al., 2019).

In the study reported in Chapter 4, I tackled the question of whether discourse particles serve as predictive cues to interpretation in context, focusing on the incremental interpretation of sentences containing the German discourse particles *eigentlich* and *tatsächlich*. I investigated the processing of referring expressions of the sort "Das ist *eigentlich/ tatsächlich* ein Bild von einem Wolf" (English: This is *eigentlich/ tatsächlich* a picture from a wolf; "That's actually/ indeed a picture of a wolf") against discourse contexts which bias not only concrete referential expectations but also clear mappings between those referential expectations and a pragmatic interpretation of the respective particles. Thus, I manipulated the predictability of any given referent as a function of an interaction between the discourse context and the usage of the particles in context. In a separate condition I also manipulated whether the particles were used reliably in context. In other words, I looked at whether pragmatically incoherent usages of *eigentlich* and *tatsächlich* lead to qualitative modulations of the processing signatures compared to the pragmatically coherent usages.

In the setup used in Chapter 4, I used a mouse-tracking task in which participants listened to a question-answer sequence where the answer consisted of a sentence containing one of

two German discourse particles, either *eigentlich* and *tatsächlich*. As discussed above, these particles index how the utterance they are embedded in relates to a contextually-relevant propositional antecedent. Importantly, this entails that the presence of either one particle in a sentence is expected to affect incremental interpretation differentially depending on the specific relation the particle signals in context. While eigentlich signals a propositional contrast, tatsächlich signals a propositional agreement. Given the setup of the experimental task, I was interested in how participants' mouse trajectories would be affected by the presence of each particle in the test sentences. More concretely, I was interested in the shape of the trajectory segments measured after the onset of the particle and before the onset of the disambiguating noun, i.e., "Das ist PARTICLE [ein Bild von einem] NOUN". Crucially, any changes in mouse position within this time window can reasonably be assumed to be an effect of processing the particle, such that movements either towards or away from the image representing the referent mentioned in the question indicate a sensitivity to the meaning of the processed particle.

The account under consideration therefore predicts that upon hearing *eigentlich* participants would shift their mouse trajectory towards the referent not mentioned in the question, as that referent constitutes a contextually-relevant alternative to the referent mentioned in the question. It also predicts that upon hearing *tatsächlich* participants would continue their trajectory towards the referent mentioned in the question. Note that, in both cases, if participants initiate any mouse movement prior to the onset of the particle, their trajectories are assumed to be en route towards the mentioned referent at the onset of the particle, as the mention of the referent in the question biases them to entertain that particular referent before any relevant semantic information from the response is processed. The results showed that, in the reliable group, both *eigentlich* and *tatsächlich* were used predictively in the interpretation of the referring expressions, such that their presence in an utterance induced participants to move their mouse towards the relevant referential target – which was different for each particle – before hearing information from the disambiguating noun.

The results also showed that exposure to unreliable particle usages not only led to temporal modulations of the qualitative patterns found in the reliable group but it also reduced the predictive potential of the particles, as evidenced by an overall reduction in the number of predictive trajectories.

I concluded from these results that discourse particles like *eigentlich* and *tatsächlich* can serve as predictive cues in the incremental processing of language, allowing comprehenders to anticipate lexical material in a sentence before actually processing that material. These particles thus function as pragmatic cues to interpretation, modulating the illocutionary force of the utterance they occur in.

1.8   Chapter 5

Much like in Chapter 4, in Chapter 5 I present a study focused on a phenomenon which can be classified as a conventional implicature. While the core underlying theoretical assumption in this chapter is therefore the same as that in Chapter 4, namely that language users reason about what the meanings conventionally associated with a particular linguistic form might mean in context, in contrast to the forms investigated in Chapter 4, which serve a specialized function as pragmatic markers, the expressions investigated in Chapter 5 have a conventionalized pragmatic meaning alongside their conventional lexical meaning. More concretely, I investigated English adjectives which have a conventionalized meaning of intensification, functioning much like a grammaticalized modifier of degree like *very* or a degree adverb like *really*. Crucially, it follows from a conventional implicature account that the pragmatic meaning of intensification conventionally associated with a given adjective is relevant for interpretation as it signals the producer's commitment to the truth of the content communicated in their utterance. As such, when interpreting an utterance with an intensifying adjective as in *He is a real weasel*, one might infer, given pragmatic expectations, that the producer believes in the truth of the statement *He is a weasel*. Given that the statement is

figurative in nature, meaning that the person being metaphorically characterized as a weasel has some quality conventionally associated in English with the animal weasel, there can be, at face value, a potential clash between the lexical meaning of *real* meaning 'reality' and its pragmatic meaning of intensification.

In the study reported in Chapter 5, I investigated the incremental interpretation of utterances containing the English intensifying adjectives *actual*, *literal*, and *real*. Previous psycholinguistic research has shown that, given general pragmatic expectations, comprehenders can use adjectives as predictive cues in the incremental interpretation of adjective-modified expressions, whereby they are able to predict referents before processing disambiguating semantic information from the nouns modified by the adjectives (e.g. D. Grodner and Sedivy, 2011; Rubio-Fernandez and Jara-Ettinger, 2020; Sedivy et al., 1999). In these studies, the adjectives provided cues to the literal meaning of the referring expressions, usually descriptions of objects of different colors and shapes. Unlike previous work focused on the processing of adjective-modified noun phrases, I investigated how comprehenders process expressions containing adjectives which function as pragmatic markers of intensification despite having a lexical semantics denoting truth or literality. What this means is that, unlike in the case of expressions modified by adjectives with a straightforward truth-conditional meaning, expressions modified by adjectives with a pragmatic meaning of intensification might be temporarily ambiguous not because they are denotationally underspecified but rather because the adjectives' pragmatic and semantic meanings might induce different, if not conflicting, interpretations during online processing.

In the setup used in Chapter 5, I used a self-paced reading task in which participants read metaphoric statements of the sort *He is a(n) ADJ weasel sometimes during games* in different experimental conditions, each condition including a different type of adjective (ADJ) before the modified noun. Crucially, by design, adjectives in the so-called *Metaphor-supporting* condition were meant to cue an interpretation which was likely metaphoric, i.e., one where a person was likely being referred to (e.g. *He is devious weasel ...*). Adjectives in

the so-called *Literal-supporting* condition, on the other hand, were meant meant to cue an interpretation which was less likely to be metaphoric, i.e., one where, in comparison to the *Metaphor-supporting* condition, it was less likely that a person was being referred to (e.g. *He is a furry weasel ...*). As discussed before, the adjective *real* in the same construction, i.e., *He is real weasel ...*, functions as a marker of intensification, one which can be construed as intensifying the meaning of *He is a weasel*. Crucial for the design was the fact that, aside from the adjective-modified noun phrase, the test sentences contained a modifier adverbial phrase headed by *sometimes*. By design, *sometimes* was meant to cue a likely metaphoric interpretation of the sentence, which, from an incremental perspective, either supports or clashes with the unfolding interpretation cued by the noun phrase, depending on whether that interpretation is more likely metaphoric, as should be the case in the *Metaphor-supporting* items (i.e. *He is a devious weasel ...*), or less likely metaphoric, as should be the case in the *Literal-supporting* items (i.e. *He is a furry weasel ...*). The key assumption underlying this setup was that, after reading the adjective and the noun, subsequent semantic material in the sentence should be more or less predictable depending on whether the adjective supports a more or less likely non-literal interpretation. The predictions related to the processing signatures measured at the adverb *sometimes*, which was the region directly following the modified noun in the critical sentence frame. This entails an effect of adjective type such that sentences with a literal-supporting adjective (*He is a furry weasel sometimes ...*) were expected to be harder to process at the adverb *sometimes* compared to sentences with a metaphor-supporting adjective (*He is a devious weasel sometimes ...*), given that the adverb provides a cue which conflicts with a literal interpretation of *He is a weasel*. Crucially, sentences with an intensifying adjective (*He is a real weasel sometimes ...*) were expected to be processed more easily at the adverb compared to sentences with a literal-supporting adjective, however, only if the intensifying adjectives were taken as pragmatic cues to non-literal meaning, i.e., if they are interpreted not in terms of their semantics related to truthfulness or reality but rather in terms of their conventionalized pragmatic meaning as intensifiers.

31

If, on the other hand, they are interpreted in terms of their truth semantics, then sentences with intensifying adjectives were expected to be processed similarly to sentences with a literal-supporting adjective.

As predicted by the account under which intensifying adjectives serve as pragmatic cues to non-literality despite their truth-conditional lexical meaning, I found that items in the *Intensifier* condition were read qualitatively similar to items in the *Metaphor-supporting* condition, such that there were no reliable mean reading time differences between the two conditions at the critical adverb region. I also found that, as predicted by the abovementioned account, items in the *Intensifier* condition were read faster than items in the *Literal-supporting* condition, such that there were reliable mean reading time differences between the two conditions at the adverb.

I concluded from these results that intensifying adjectives can act as pragmatic cues in the incremental interpretation of language. More concretely, they can serve as *predictive* cues to non-literal meaning, such that in the processing of non-literal, metaphoric statements they cue a likely non-literal interpretation despite their lexical semantics denoting truth or reality. This is in line with the hypothesis that, given their function as pragmatic markers of intensification, *actual*, *literal*, and *real* are more likely to be interpreted in terms of their intensificational meaning than in terms of their lexical semantic meaning in situations in which the former supports the unfolding interpretation more strongly than the latter.

REFERENCES

Ariel, Mira (2016). "Revisiting the typology of pragmatic interpretations". In: *Intercultural Pragmatics* 13.1, pp. 1–35.

Bergen, Leon and Daniel J Grodner (2012). "Speaker knowledge influences the comprehension of pragmatic inferences." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 38.5, p. 1450.

Blakemore, Diane (1987). "Semantic constraints on relevance". In:

Bubić, Andreja, D. Yves von Cramon, and Ricarda I. Schubotz (2010). "Prediction, Cognition and the Brain". In: *Frontiers in Human Neuroscience* 4.

Clark, Andy (2013). "Whatever next? Predictive brains, situated agents, and the future of cognitive science." In: *The Behavioral and brain sciences* 36 3, pp. 181–204.

Cornillie, Bert (2009). "Evidentiality and epistemic modality: On the close relationship between two different categories". In: *Functions of language* 16.1, pp. 44–62.

Dänzer, Lars (2020). "The explanatory project of Gricean pragmatics". In: *Mind & Language.*

de Brabanter, Philippe and Patrick Dendale (2008). "Commitment: the term and the notions". In: *Belgian Journal of Linguistics* 22.1, pp. 1–14.

de Saussure, Louis and Peter Schulz (2007). "Interfacing pragmatics". In: *Pragmatics and Cognition* 15.1, p. 3.

Degen, Judith and Michael K Tanenhaus (2016). "Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study". In: *Cognitive science* 40.1, pp. 172–201.

— (2019). "Constraint-based pragmatic processing". In: *The Oxford handbook of experimental semantics and pragmatics*, pp. 21–38.

Diewald, Gabriele (2011). "Pragmaticalization (defined) as grammaticalization of discourse functions". In: *Linguistics* 49.2, pp. 365–390. DOI: doi:10.1515/ling.2011.011.

Domaneschi, Filippo, Maribel Romero, and Bettina Braun (2017). "Bias in polar questions: Evidence from English and German production experiments". In: *Glossa: a journal of general linguistics* 2.1.

Feng, Guangwu (2011). "A neo-Gricean pragmatic analysis of Chinese pragmatic markers". In: *Language Sciences* 33.3, pp. 417–434. DOI: https://doi.org/10.1016/j.langsci.2010.10.005.

— (2010). *A Theory of Conventional Implicature and Pragmatic Markers in Chinese*. Leiden, The Netherlands: Brill. DOI: https://doi.org/10.1163/9789004253179.

Ferreira, Fernanda and Suphasiree Chantavarin (2018). "Integration and prediction in language processing: A synthesis of old and new". In: *Current directions in psychological science* 27.6, pp. 443–448.

Franke, Michael (2016). "Task types, link functions & probabilistic modeling in experimental pragmatics". In: *Pre-proceedings of "Trends in Experimental Pragmatics"*, pp. 56–63.

Geurts, Bart (2016). "A wish list for experimental pragmatics". In: *Pre-proceedings of "Trends in Experimental Pragmatics"*, pp. 68–70.

— (2019). "What's wrong with Gricean pragmatics". In: *ExLing 2019: Proceedings of 10th International Conference of Experimental Linguistics,*

Gibbs, Raymond W. and Herbert L. Colston (2020). "Pragmatics Always Matters: An Expanded Vision of Experimental Pragmatics". In: *Frontiers in Psychology* 11.

Goldrick, Matthew (2022). "An Impoverished Epistemology Holds Back Cognitive Science Research". In: *Cognitive Science* 46.9, e13199.

Gotzner, Nicole and Jacopo Romoli (2022). "Meaning and Alternatives". In: *Annual Review of Linguistics* 8.1, pp. 213–234.

Gray, Bethany and Douglas Biber (2012). "Current conceptions of stance". In: *Stance and voice in written academic genres*. Springer, pp. 15–33.

Grice, H. Paul (1975). "Logic and conversation". In: *Syntax and Semantics* 3, pp. 41–58.

Grodner, Daniel and Julie C Sedivy (2011). "10 The Effect of Speaker-Specific Information on Pragmatic Inferences". In: *The processing and acquisition of reference*. Vol. 2327. MIT Press, pp. 239–272.

Gumperz, John J and Stephen C Levinson (1996). "Introduction: Linguistic relativity reexamined". In: *Rethinking linguistic relativity*. Cambridge University Press, pp. 1–20.

Gutzmann, Daniel and Hans-Martin Gärtner (2013). *Beyond Expressives: Explorations in Use-Conditional Meaning*. Leiden, The Netherlands: Brill. DOI: https://doi.org/10.1163/9789004183988.

Gutzmann, Daniel and Katharina Turgay (2019). "Secondary Content: An Introduction". In: *Secondary Content: The Semantics and Pragmatics of Side Issues*. Leiden, The Netherlands: Brill, pp. 1–25. DOI: https://doi.org/10.1163/9789004393127_002.

Huang, Yi Ting and Jesse Snedeker (2018). "Some inferences still take time: Prosody, predictability, and the speed of scalar implicatures". In: *Cognitive psychology* 102, pp. 105–126.

Huettig, Falk and Nivedita Mani (2016). "Is prediction necessary to understand language? Probably not". In: *Language, Cognition and Neuroscience* 31.1, pp. 19–31.

Ilie, Cornelia (2002). "Who's afraid of Paul Grice? The role of the cooperative principle in academic metadiscourse". In: *RASK* 16, pp. 3–32.

Kamide, Yuki, Gerry TM Altmann, and Sarah L Haywood (2003). "The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements". In: *Journal of Memory and language* 49.1, pp. 133–156.

Kuperberg, Gina R and T Florian Jaeger (2016). "What do we mean by prediction in language comprehension?" In: *Language, cognition and neuroscience* 31.1, pp. 32–59.

Kutas, Marta, Katherine A DeLong, and Nathaniel J Smith (2011). "A look around at what lies ahead: Prediction and predictability in language processing." In:

Litwin, Piotr and Marcin Miłkowski (2020). "Unification by fiat: arrested development of predictive processing". In: *Cognitive Science* 44.7, e12867.

Nieuwland, Mante S., Tali Ditman, and Gina R. Kuperberg (2010). "On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities". In: *Journal of memory and language* 63.3, pp. 324–346.

Noveck, Ira (2018). *Experimental pragmatics: The making of a cognitive science*. Cambridge University Press.

Potts, Christopher (2007). "Into the Conventional-Implicature Dimension". In: *Philosophy Compass* 2, pp. 665–679.

Rasenberg, Marlou, Joost Rommers, and Geertje van Bergen (2019). "Anticipating predictability: an ERP investigation of expectation-managing discourse markers in dialogue comprehension". In: *Language, Cognition and Neuroscience* 35, pp. 1–16.

Ricci, Claudia and Corinne Rossari (2018). *Commitment phenomena through the study of evidential markers in Romance languages*.

Ridderinkhof, K Richard (2017). "Emotion in action: A predictive processing perspective and theoretical synthesis". In: *Emotion Review* 9.4, pp. 319–325.

Rohde, Hannah and Chigusa Kurumada (2018). "Alternatives and inferences in the communication of meaning". In: *Psychology of Learning and Motivation*. Vol. 68. Elsevier, pp. 215–261.

Rubio-Fernandez, Paula and Julian Jara-Ettinger (2020). "Incrementality and efficiency shape pragmatics across languages". In: *Proceedings of the National Academy of Sciences* 117.24, pp. 13399–13404.

Sander, Thorsten (2022). "Taxonomizing Non-at-Issue Contents". In: *Grazer Philosophische Studien*.

Sedivy, Julie C et al. (1999). "Achieving incremental semantic interpretation through contextual representation". In: *Cognition* 71.2, pp. 109–147.

Sperber, Dan and Deirdre Wilson (1986). "Relevance: Communication and Cognition". In:

Tonhauser, Judith et al. (2013). "Toward a Taxonomy of Projective Content". In: *Language* 89, pp. 109–66.

Traugott, Elizabeth Closs (2012). "Pragmatics and language change". In: *The Cambridge handbook of pragmatics*, pp. 549–566.

— (1995). "Subjectification in grammaticalization". In: *Subjectivity and subjectivisation: Linguistic perspectives* 1, pp. 31–54.

Urbach, Thomas P, Katherine A DeLong, and Marta Kutas (2015). "Quantifiers are incrementally interpreted in context, more than less". In: *Journal of memory and language* 83, pp. 79–96.

Bergen, Geertje van and Hans Rutger Bosker (2018). "Linguistic expectation management in online discourse processing: An investigation of Dutch inderdaad 'indeed' and eigenlijk 'actually'". In: *Journal of Memory and Language*.

Wilson, Deirdre (2019). "Relevance Theory". In: *Oxford Research Encyclopedia of Linguistics*.

Winkler, István, Susan L Denham, and Israel Nelken (2009). "Modeling the auditory scene: predictive regularity representations and perceptual objects". In: *Trends in cognitive sciences* 13.12, pp. 532–540.

Chapter Two

Interpreting utterances with scalar quantifiers and tracking referents in the online processing of referring expressions

2.1 Introduction

In this chapter, the first empirical chapter of the dissertation, I present a case study focusing on a well-studied – and arguably well understood – family of pragmatic phenomena, namely scalar inferences, which serve as a window into issues concerning prediction and predictability in pragmatic processing. From a theoretical perspective, the derivation of an inference cued by a scalar expression rests on the assumption introduced in Chapter 1 that language users reason about the informativity of linguistic alternatives. More specifically, the pragmatic derivation is assumed to be linked to the informativity of lexical alternatives which are construed as being ordered on a scale according to their relative semantic strength[1]. Crucially,

---

[1]While it is debated what exact form of inferencing best characterizes the derivation of scalar implicatures (see e.g. Gotzner and Romoli, 2022), the current empirical picture suggests that lexical scales play an important role in the interpretation of scalar terms, therefore supporting a characterization of scalar inferences as generalized conversational implicatures dependent primarily on the conventionalized (semantic) meaning of scalars as opposed to *ad hoc* inferencing in context, which would instead support a characterization in

it follows from this account that the relationship between alternatives on a particular lexical scale – as between the quantifiers *some* and *all* on a scale of quantity – is pragmatically relevant for interpretation as different items on the scale are more or less informative in context depending on their semantic status relative to one another. As such, when interpreting an utterance with a semantically weak scalar, one might infer, given (Gricean) pragmatic expectations, that the expression is being used informatively and thus that it's the strongest possible scalar alternative in that given situation. For instance, when interpreting an utterance with a scalar expression like the quantifier *some* one might therefore infer that, unless suggested otherwise by other cues in the signal or in the discourse context, the set of interest is that of some but likely not all, given that the stronger alternative *all* was not actually uttered.

There is a vast experimental literature on the topic of pragmatic enrichment of scalars, with plenty of experimental work dedicated to the analysis of implicature triggering and derivation (see Gotzner and Romoli (2022) for a recent overview). Given the breadth of psycholinguistic work on the processing of scalars, I focus my attention on a particular linguistic form which has been classically associated with scalar inferences: the quantifier *some*, or rather, in my case, its German counterpart *einige*. Despite previous research having shed light on the circumstances under which processing *some* might give rise to an enriched pragmatic interpretation, the relative predictability of either an enriched or a non-enriched interpretation of a sentence containing a lower-bounded scalar like *some* has not been directly investigated. In the study reported in the present chapter, I looked exactly at the *relative* predictability of subsequent material in a sentence when processing the lower-bounded scalar *einige* – which might give rise to an implicature given its informativity in context – as opposed to the upper-bounded scalar *alle* (English: *all*), which might serve as a predictive cue to meaning without giving rise to a pragmatic inference. Before introducing the experiment and reporting my findings in detail, I review some of the previous findings

---

terms of particularized implicatures/ explicatures.

on prediction in pragmatic processing, in particular those related to the processing of scalar quantifiers.

2.2   Pragmatic prediction in online processing

As discussed in Chapter 1, prediction in online processing has been a central theme in recent cognitive scientific and psycholinguistic research (see, inter alia, Bubić et al., 2010; Clark, 2013; Kuperberg and Jaeger, 2016). Numerous studies have shown that humans are able to predict upcoming input on the basis of perceptual, social, and linguistic cues (e.g. Kamide et al., 2003; Ridderinkhof, 2017; Winkler et al., 2009; see also Litwin and Miłkowski, 2020). As far as linguistic prediction is concerned, studies have shown that material can be pre-activated at different levels of linguistic representation, from phonologically- and lexically-driven pre-activation to pre-activation derived from syntactic and semantic cues (e.g. Boudewyn et al., 2015; Kuperberg and Jaeger, 2016; Mani and Huettig, 2012; Urbach et al., 2015), although some of these findings – particularly with respect to phonological prediction – have failed to replicate in more recent studies (e.g. Nieuwland, Arkhipova, et al. 2020; see also Kuperberg and Jaeger (2016) for a detailed discussion of the different sorts of prediction potentially involved in predictive language processing).

As for pragmatics, previous psycholinguistic studies have provided evidence that high-level semantic and pragmatic prediction can occur while people process language, from the processing of negation (e.g. Haase et al., 2019; Nieuwland, 2016; see also Scappini et al., 2015) to the processing of other pragmatically relevant linguistic cues such as scalar quanti-fiers (e.g. Augurzky et al., 2019; Nieuwland, Ditman, et al., 2010). Although the bulk of the evidence stems from neurolinguistic studies, other studies have drawn on behavioral methods such as eye-tracking (e.g. Degen and Tanenhaus, 2016; Huang and Snedeker, 2018; Scholman et al., 2016) and self-paced reading (e.g. Bergen and Grodner, 2012; Bicknell and Rohde, 2009). For example, studies relying on the manipulation of the pictorial context which the

linguistic stimuli refer to show that people are sensitive to contextually induced pragmatic expectations (Augurzky et al., 2019; Darley et al., 2020; Degen and Tanenhaus, 2016; Spychalska et al., 2016). In other words, there is evidence that language users form expectations about an unfolding linguistic signal based on expectations of how pragmatically felicitous an utterance might be in a particular discourse context, including visually-anchored contexts.

In the present study, I employed an experimental setup which involved manipulating the visual context an utterance refers to. While much previous work on pragmatic processing in visually-anchored contexts has focused on EEG studies, the study below investigates whether visually-anchored pragmatic expectations also affect self-paced reading times. I used referring expressions containing scalar quantifiers, which are known to act as pragmatic cues in the online processing of language. In previous research, predicatibility has been linked to the relative probability of a word serving as a continuation of an unfolding linguistic signal (Kuperberg and Jaeger, 2016; Levy, 2008). In the particular case of written language comprehension, the surprisal of a given word as a sentence continuation has been argued to be proportional to the cognitive effort of reading a word (Smith and Levy, 2013). Reading times (RTs) are thus assumed to be proportional to the effort involved in processing words in a written stream of language. Following the previous literature on visually-anchored pragmatic processing, I address *pragmatic surprisal theory*, which states that, at the level of utterance and discourse processing, language users attend to different magnitudes of pragmatic felicity and that the resulting pragmatic expectations lead to slower reading times for words with higher contextually-anchored pragmatic surprisal. To carve out concrete qualitative predictions from pragmatic surprisal theory, I assume that comprehenders are sensitive, at the very least, to the contrast between semantically congruent and semantically incongruent utterances, as well as to the contrast between pragmatically felicitous and pragmatically infelicitous utterances.

To recapitulate the basic assumptions, in a Gricean framework, the pragmatic import of an utterance can be analyzed in terms of considerations of quantity, a notion which is usu-

ally equated with informativeness, where a form is said to render an utterance either more or less informative depending, for instance, on other semantically-related alternatives which together with the uttered form constitute a salient linguistic scale (Horn, 1989). While there are other aspects of pragmatic felicity which may be of relevance to the processing of utterances containing scalars, in the study reported below[2] I operationalize pragmatic felicity as informativity in the Gricean sense. This implies treating underinformative utterances as pragmatically infelicitous, such that, in the case of a referring expression, an underinformative utterance is a true description for which a salient alternative exists which is also true and logically stronger, so that the latter entails the former by semantic meaning but not the other way around. To illustrate how pragmatic felicity in this sense can lead to concrete predictions about expectation-based processing difficulty, consider Figure 2.1 (which shows example items from the experiments presented in section 2.3 and 2.4 below).
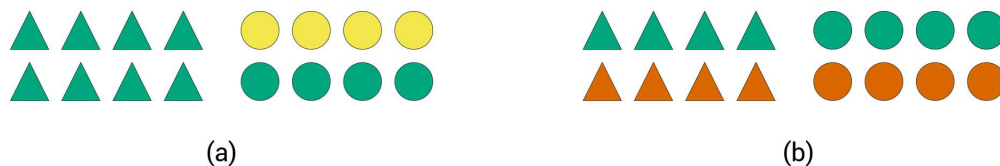
Figure 2.1 Examples of visually presented context information which raises expectations of particular lexical material based on considerations of pragmatic felicity.

In the terms discussed above, an unfolding linguistic signal such as *Some of the ...* may be said to be an informative description with reference to scene (a). Indeed, even though the referring expression is in principle semantically congruent with either shape array depicted in the scene, from a pragmatic perspective one can expect such a description to eventually refer to the circle array and not the triangle array, given that a salient alternative expression, namely *All of the ...*, would have been semantically stronger and thus pragmatically more informative in case the triangle array was the actual target of the referring expression. This

---

[2]This study has been conceived and carried out in collaboration with Michael Franke.

inferential jump from a lower-bounded interpretation of *some* to an upper-bounded one is regarded as a pragmatic enrichment, the so-called scalar implicature, whereby an interpreter is able to derive the enriched, upper-bounded meaning of the referring expression on the basis of an assumed informativity relation between the scalar alternatives *some* and *all*, which, couched in a more general expectation of pragmatic felicity, gives rise to probabilistic expectations about the to-be-mentioned shape array. Importantly, notice that a similar inference would not be possible with reference to scene (b), where neither of the shape arrays can be truthfully described using the quantifier *all*, hence there being no stronger alternatives to *some*.

Now, considering that people generate online predictions based on pragmatic expectations, their predictions should vary as a function of both the amount of linguistic information available at any given moment, i.e., how much of the unfolding linguistic signal has been processed, as well as one's overall expectations regarding what is being communicated, in this case what is being referred to. Crucially, the two are interrelated, such that expectations might shift as new material is processed and integrated into competing sentence and discourse models. In practical terms, one needs to establish a link hypothesis between the predictions derived from a high-level pragmatic theory and the empirically measured processing signatures which are to reflect the effects of the postulated underlying pragmatic mechanisms. I do so in the present case via surprisal, such that I expect that comprehenders will read a given description more slowly whenever their expectations fail to be met by whatever fragment of the description they are reading. That is, if a participant expects to read a specific word at a specific moment but instead reads a different word, they are expected to experience processing difficulties as a result of a mismatch between their expectations and the actual linguistic material they encounter. This is in line with previous work which has established surprisal as a possible link hypothesis in language processing, including processing as measured by means of online reading times (e.g., Demberg and Keller, 2008; Monsalve et al., 2012; Smith and Levy, 2013).

Bringing together the theoretically-motivated considerations of pragmatic interpretation with the insights about the integration of linguistic cues in online processing via surprisal, one arrives at the following expected effects on RTs, referred henceforth as *pragmatic surprisal theory* (PST): after having read a sentence initial fragment of words $w_1, \ldots w_n$, comprehenders read word w_{n+1} faster in (visual) context C than in context C' if they expect w_{n+1} with a higher probability to occur in C than in C'. Given a quantified expression like *Some of the triangles are green*, I hypothesize that, after having read the quantifier, a comprehender will read the next critical term, i.e., the shape term, more slowly in context (b) than in context (a). This is because after processing the quantifier, more than one true and pragmatically felicitous continuation is available in (b), whereas in (a) only one true and pragmatically felicitous continuation is available (e.g., Augurzky et al., 2019; Spychalska et al., 2016). Therefore, in a scenario like (a) comprehenders are expected to generate strong predictions about a specific sentence continuation, which implies strong expectations about the to-be-read shape term in the unfolding sentence. In a scenario like (b), however, comprehenders are expected to generate weaker predictions about one or the other possible sentence continuations, which implies weaker relative expectations about the to-be-read shape term in the unfolding sentence. Whether a given prediction A or B is confirmed, participants are expected to be surprised by what they read, given that they did not have strong expectations about a given term X or Y in the first place. In short, given the cases illustrated in Fig. 2.1, a comprehender confronted with (a) should expect to read *circles* after reading *Some of the*, as that is the most informative continuation at that point in the sentence. Crucial for my argumentation, however, this prediction should only hold when the scalar term *some* is enriched pragmatically, giving rise to a so-called scalar implicature, as discussed above.

The remainder of the chapter is structured as follows. Section 2.3 introduces the experimental material and the first part of the experiment, a sentence completion task aimed to obtain information about which descriptions participants themselves would generate for the pictorial materials. This data serves to ground our specific assumptions about which kinds of

pragmatic expectations participants may have during processing. Section 2.4 focuses on the second part of the experiment, a self-paced reading task, describing the design and discussing the results. Section 2.5 summarizes the findings and concludes.

2.3 Task 1 - Production of referring expressions containing scalar quantifiers

In the main task, described below in section 2.3.3, I tested the extent to which comprehenders process referring expressions containing scalar terms predictively on the basis of pragmatic considerations. My working assumptions are derived from general, Gricean-inspired considerations of pragmatic felicity and informativity. More specifically, I assume that comprehenders prefer true and pragmatically felicitous utterances over true but pragmatically infelicitous utterances. However, in practice, comprehenders might have somewhat different expectations, expectations that perhaps diverge from or fuse together these different theoretically-motivated assumptions. In order to systematically tease apart pragmatic surprisal from other possible auxiliary assumptions, participants were first asked to perform a production task in order to determine whether their observed production behavior gives rise to expectations that support a surprisal-based account of the reading data reported in section 2.3.3. My aim here was to test the extent to which comprehenders' empirically verified reading patterns, assumed to be linked to their underlying pragmatic expectations, match those derived from a normative, theoretically-grounded account of pragmatic processing.

Before being asked to perform the reading task, participants were asked to describe the same stimuli used in that task by completing sentences which had gaps in them. As explained earlier in the introduction, the aim of such task was to collect descriptions of the stimuli so as to know what the likelihood of producing specific descriptions might be in the first place. I was thus particularly interested in knowing what possible sentence configurations are more

likely given the specific types of pictures found in our stimulus set.

2.3.1 Method

Participants

Fifty-eight native speakers of German were recruited among the cognitive science and psychology student population of the University of Osnabrück. Participants were given course credit in exchange for their participation in the experiment. Data collection was conducted at the Institute of Cognitive Science of the same university, in a computer laboratory designed for the execution of behavioral experiments.

Materials and design

Participants were asked to complete 21 German sentences, each referring to a visual scene composed of two arrays of eight geometric shapes. Figure 2.2 exemplifies the different types of pictures contained in the study, which varied along three semantic dimensions: the color of the shapes (Fig. 2.2, first row), their size (Fig. 2.2, second row), and their position relative to a box also depicted in a subset of the scenes (Fig. 2.2, third row).
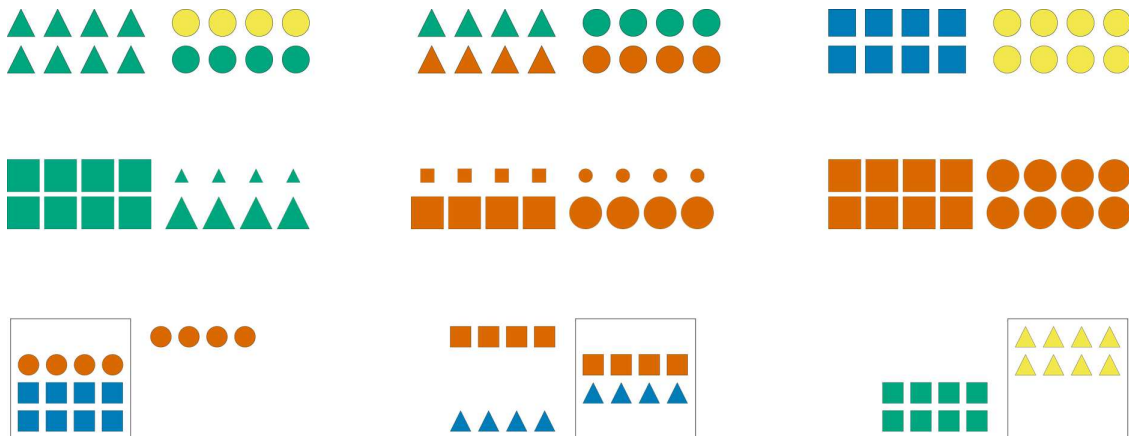
Figure 2.2 Sample visual scenes. Scenes in the first column contain one homogeneous shape array and one heterogeneous shape array; scenes in the second column contain two heterogeneous shape arrays; scenes in the third column contain two homogeneous shape arrays.

Three choices were recorded per sentence, each of these mapping onto the critical regions from the reading task, as in the sentence structure below:

Original: QUANT | der | SHAPE | auf | dem | Bild | sind | PROP | [der | Box].

English gloss: QUANT | of the | SHAPE | in | the | picture | are | PROP | [the | box].

Each trial, participants' task was to select one scalar quantifier (QUANT), one shape term (SHAPE), and one adjective or preposition, which depending on the picture type related to a different visuo-semantic property (PROP). There were two quantifiers to choose from, three shape terms, two prepositions, and six adjectives, as shown below.

- QUANT – *einige* (some), *alle* (all);

- SHAPE – *Kreise* (circles), *Quadrate* (squares), *Dreiecke* (triangles);

- PROP – *orange* (orange), *blau* (blue), *gelb* (yellow), *grün* (green), *klein* (small), *groß* (big), *in* (inside), *neben* (next to).

Procedure

Written as well as oral instructions were provided prior to the actual task, followed by three practice trials which mimicked the exact procedure of the test trials. Each trial, participants were presented with a visual scene as well as with sentence with gaps in it. Participants were instructed to fill the gaps in the sentence by choosing between different words in drop-down menus available on screen. After making their choices and filling all gaps, a 'next' button appeared on screen, allowing participants to proceed to the next trial.

2.3.2   Results and discussion

The data and analysis scripts are available for inspection under https://zenodo.org/record/5156186#.Y1kywYTMKUk. Recall that for every test sentence participants chose a quan-

tifier, a shape term, as well as a third term which, depending on the type of picture, could either be an adjective denoting color or size or a preposition. Figure 2.3 shows the proportion of produced combinations of quantifier + shape term + property term, anchored to sample items where color is the relevant property to be described. Each row shows the production preferences for a given type of visual scene, the left column showing the choices of expressions containing the quantifier *alle*, and the right column showing the choices of expressions containing the quantifier *einige*. The color of the bars represent the respective color term used in the descriptions.
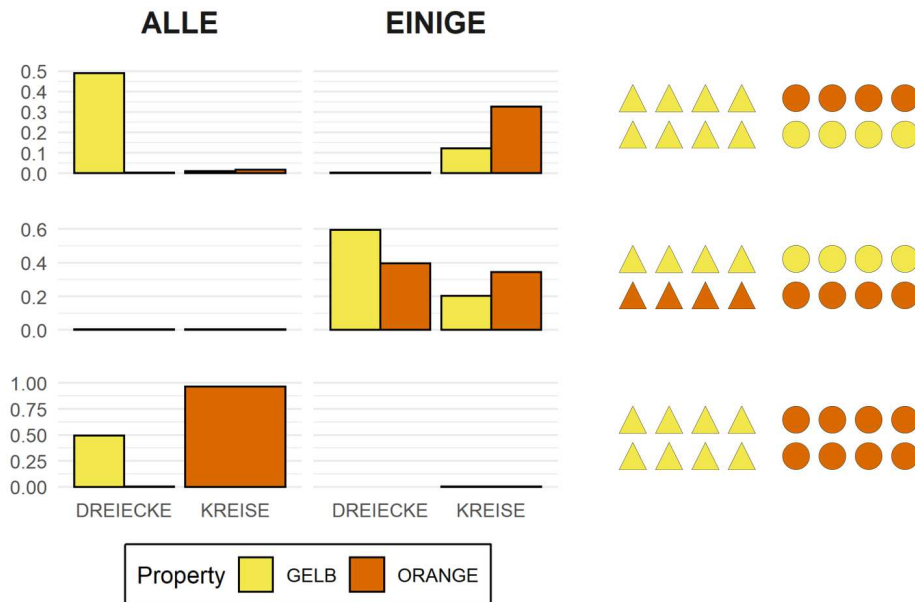
Figure 2.3 Proportion of production choices by picture type. The written labels indicate the type of expression selected by participants, which included a quantifier (*alle* vs. *einige*), a shape term (*Dreiecke* vs. *Kreise*), and a property term (*gelb* vs. *orange*). The left column shows descriptions containing the quantifier *alle*, while the right column shows descriptions containing the quantifier *einige*.

Unsurprisingly, the results show that participants produced descriptions of pictures with two homogeneous arrays [bottom row] virtually always using the quantifier *alle* (left column), much the same way that they produced descriptions of pictures with two heterogeneous

arrays [middle row] virtually always using the quantifier *einige* (right column). This is in line with the naive assumption that people prefer semantically congruent utterances over semantically incongruent ones. The results also show that, when describing pictures with only one heterogeneous array [top row], participants prefer producing descriptions containing the quantifier *alle*, that is, they prefer referring to the homogeneous array in the scene. This can be interpreted as a bias for producing descriptions using terms which are semantically stronger, and thus pragmatically more informative, than their scalar alternatives.

Another bias participants show when describing pictures with only one heterogeneous array is to prefer referring to the semantic dimension which is congruent exclusively with the heterogeneous array. In other words, when people use *einige* to refer to a scene like the one depicted on the top row of Fig. 2.3, they tend to couple the quantifier with a property term that renders the description maximally distinct from other possible descriptions of that same picture, such as *Einige der Kreise auf dem Bild sind* <u>*orange*</u> (English: Some of the circles in the picture are <u>orange</u>). Both this preference as well as the preference for semantically strong lexical alternatives are in line with the naive assumption that, all things being equal, people prefer pragmatically felicitous utterances over pragmatically infelicitous ones. Yet, despite these preferences, the question remains as to whether the normative expectations supported by such production data are borne out in the processing data.

2.4 Task 2 - Processing of referring expressions containing scalar quantifiers

2.4.1 Method

Materials and design

Participants read 84 German sentences, each referring to a varying visual scene, as per the stimuli in Task 1. Participants saw four instances of each picture type for each condition,

for a total of 84 trials. Reading times were measured at eight to ten sentence regions, each of these consisting of one word of the sentence structure introduced above:

QUANT | der | SHAPE | auf | dem | Bild | sind | PROPER | [der | Box].

In the case of the reading task, the terms represented here by uppercase words varied each trial. Much like in Task 1, each sentence contained a scalar quantifier (QUANT), a shape term (SHAPE), and either an adjective denoting color or size or a preposition (PROPER):

- QUANT – *einige* (some), *alle* (all);

- SHAPE – *Kreise* (circles), *Quadrate* (squares), *Dreiecke* (triangles);

- PROPER – *orange* (orange), *blau* (blue), *gelb* (yellow), *grün* (green), *klein* (small), *groß* (big), *in* (inside), *neben* (next to).

Different sentences and pictures were paired so as to yield four critical experimental conditions. These conditions differ in terms of the quantifier which was processed as well as in terms of whether the matching visual scene induced specific linguistic expectations at the SHAPE region. In other words, the experimental manipulation involved modulating the predictability of the shape terms by means of varying the visual context which participants encountered immediately before reading the test sentence. In this setup, predictability should, therefore, be primarily dependent on Gricean-like informativity considerations, as described above, such that readers are expected to find a given critical term more or less predictable given the assumption that a referring expression produced by a cooperative describer is informative. As such, the resulting experimental conditions are as follows:

- Alle (Biased) - Sentences containing the quantifier *alle* were paired with visual scenes which were meant to increase the predictability of specific critical terms (i.e. one of the arrays is homogeneous while the other is heterogeneous - the contrast should bias the expectation of the critical terms which match the homogeneous array). A sample pairing is as follows:
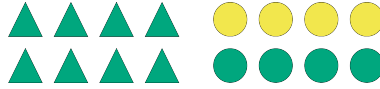
Original: "Alle der Dreiecke auf dem Bild sind grün."

English gloss: "All of the triangles in the picture are green."

Upon reading *alle*, an interpreter who is expecting a semantically congruent utterance should predict the description to refer to the triangle array, and not the circle array, as only in the case of the former is the description congruent with the scene.

- Alle (Unbiased) - Sentences containing the quantifier *alle* were paired with visual scenes which were meant not to increase the predictability of any subsequent critical term (i.e. both arrays are homogeneous). A sample pairing is as follows:
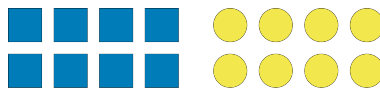
Original: "Alle der Kreise auf dem Bild sind gelb."

English gloss: "All of the circles in the picture are yellow."

- Einige (Biased) - Sentences containing the quantifier *einige* were paired with visual scenes which were meant to increase the predictability of specific critical terms (i.e. one of the arrays is heterogeneous while the other is homogeneous - the contrast should bias the expectation of the critical terms which match the heterogenous array, however, only if *einige* is read as *some-but-not-all*). A sample pairing is as follows:
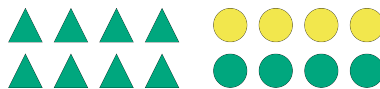
Original: "Einige der Kreise auf dem Bild sind gelb."

English gloss: "Some of the circles in the picture are yellow."

While in the case of the Alle (Biased) condition the sentence bias originates from an expectation of truthfulness, in the case of Einige (Biased) the bias originates from an expectation of informativeness, which, according to the inferential account I'm entertaining, should give rise to an implicature.

- Einige (Unbiased) - Sentences containing the quantifier *einige* were paired with visual scenes which were meant not to increase the predictability of any subsequent critical term (i.e. both arrays are heterogeneous). A sample pairing is as follows:
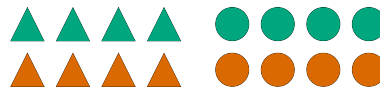
Original: "Einige der Kreise auf dem Bild sind orange."

English gloss: "Some of the circles in the picture are orange."

In addition to the four critical conditions, the experiment included three conditions which served as a baseline for the manipulations involving pragmatic expectations. These consist of pairings of sentences and pictures which resulted in semantically incongruent descriptions, as well as a condition in which descriptions containing the quantifier *einige* are semantically congruent but underspecified:

- Alle (False) - Sentences containing the quantifier *alle* were paired with visual scenes which are semantically incongruent with the referring expression (i.e. the referred array is not homogeneous). A sample pairing is as follows:
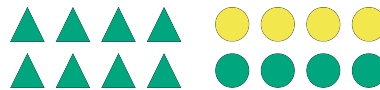
Original: "Alle der Kreise auf dem Bild sind gelb."

English gloss: "All of the circles in the picture are yellow."

51

- Einige (False) - Sentences containing the quantifier *einige* were paired with visual scenes which are semantically incongruent with the referring expression (i.e. both arrays are heterogeneous but the shape term renders the description semantically incongruent). A sample pairing is as follows:
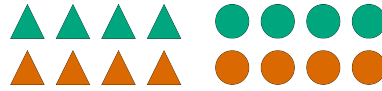
"Einige der Quadrate auf dem Bild sind orange."

- Einige (Infelicitous) - Sentences containing the quantifier *einige* were paired with visual scenes which were meant to increase the predictability of specific critical terms (i.e. one of the arrays is heterogeneous while the other is homogeneous - the contrast should bias the expectation of the critical terms which match the heterogenous array). A sample pairing is as follows:
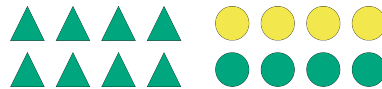
"Einige der Dreiecke auf dem Bild sind grün."

English gloss: "Some of the triangles in the picture are green."

Note that in this condition the description, which contains the quantifier *einige*, refers to the homogeneous array, which implies that if the utterance is ultimately interpreted as a congruent description of the scene no pragmatic enrichment must have taken place, suggesting a strictly semantic interpretation of the quantifier *einige*, meaning *some-and-maybe-all*.

In summary, the study contained seven experimental conditions - four critical conditions and three control conditions - as schematized in Table 2.1.

	Quantifier	
	Alle	*Einige*
Bias	Alle (Biased)	Einige (Biased)
	Alle (Unbiased)	Einige (Unbiased)
Control	Alle (False)	Einige (False)
		Einige (Infelict)

Table 2.1 Study design.

Procedure

Written as well as oral instructions were provided to participants prior to the task, followed by three practice trials which mimicked the exact procedure of the test trials[3]. All trial elements flashed in and out of the screen in a pre-specified order. First, a fixation cross appeared at the center of the screen. After 500 milliseconds, the cross disappeared and the visual scene became visible at the cross location. The picture remained on screen for as long as participants wished, and it was dismissed by a press of the space bar. Immediately after the picture disappeared, underscores appeared below the picture location, along with the written indication "Press the SPACE bar to reveal the words", presented in German. As indicated by the cue, each press of the space bar revealed one of the sentence chunks foreshadowed by the underscores. Once participants reached the last chunk in the sentence,

---

[3]The experimental program is available for inspection under https://spr-in-lab-june-2019-direct-folder.netlify.com/.

their next key press triggered the question "How accurate was the sentence as a description of the picture?", presented in German. As indicated by the question, participants' task was to rate, on a 7-point scale, how appropriate the sentence was as a description of the visual display they saw on screen. After making their choice, participants were forwarded to the next trial.

There were 14 trials of each trial type and all participants saw all of these, for a total of 84 trials. Given the constraints imposed by the experimental manipulations, similar or even identical visual scenes were paired with different referring expressions. However, the matching between an image and a description always resulted in unique trial instantiations. The 84 trials were administered in four blocks of 21 trials. In between blocks, participants encountered a pause screen, and they were encouraged to take as much time as needed before proceeding to the next block.

2.4.2 Hypotheses

Given the design of the study and the considerations outlined in the introduction, I put forward the following general prediction: participants will read the critical terms more slowly if they are unexpected. Thus, I hypothesize that at the shape region, i.e., the next critical region after the quantifier, participants will read the critical term more slowly in the unbiased conditions than in the biased conditions, regardless of the quantifier they encounter. This is under the assumption that, after processing the quantifier, more than one informative continuation is available per scene in the unbiased scenarios, whereas in the biased scenarios only one informative continuation is available. Therefore, participants in the biased conditions are expected to generate strong predictions about a specific sentence continuation, which implies strong expectations about the to-be-read shape term in the unfolding sentence. In the unbiased conditions, however, participants are expected to generate weaker predictions about one or the other possible sentence continuations, which implies weaker

relative expectations about the to-be-read shape term in the unfolding sentence. Whether a given prediction A or B is confirmed, participants are expected to be surprised by what they read, given that they did not have strong expectations about a given term X or Y in the first place. Fig. 2.4 shows what a biased and an unbiased scenario look like in relation to a description containing the quantifier *einige*.

(1) Biased visual scene   **(2)** Unbiased visual scene

Figure 2.4 Biased and unbiased scenarios for descriptions containing the quantifier *einige*.

Considering the cases illustrated in Fig. 2.4, a participant confronted with (a) should expect *Kreise* (English: circles) after reading *Einige der* (English: Some of the), as that is the most informative continuation at that point in the sentence. This prediction should only hold, however, when the quantifier *einige* is taken as a pragmatic, as opposed to a semantic, cue. Thus, I expect that the bias manipulation will yield different results depending on whether participants process the sentences as pragmatically enriched or strictly semantic descriptions. Namely, I expect that when *einige* is interpreted as *some-and-maybe-all* reading times will be similar in both the biased and unbiased conditions, as in both cases participants are expected not to generate strong predictions about one or the other possible sentence continuations.

Pragmatic interpretation	Semantic interpretation
(a) Einige | der | Kreise \ Dreiecke ...	(a) Einige | der | Kreise \ Dreiecke ...
(b) Einige | der | Kreise \ Dreiecke ...	(b) Einige | der | Kreise \ Dreiecke ...

I assume that, after processing the quantifier and the shape term, participants will then generate predictions about the last semantically relevant term in the sentence. Considering again the cases in Fig. 2.4, one may assume that in the unbiased scenario, even if participants

already know the shape term, it is still unclear which color might be referred to. Much the same way, in the biased scenario, even if participants already know that the sentence refers to circles, the color of the referred array cannot be easily predicted. In fact, even if *einige* is taken as a pragmatic cue (*some-but-not-all*), participants should not be able to generate strong predictions about a particular color term.

Pragmatic interpretation	Semantic interpretation
(a) Einige | der | Kreise | auf | dem | Bild | sind | gelb \ grün	(a) Einige | der | Kreise | auf | dem | Bild | sind | gelb \ grün
(b) Einige | der | Kreise | auf | dem | Bild | sind | grün \ orange	(b) Einige | der | Kreise | auf | dem | Bild | sind | grün \ orange

In summary, then, the following predictions are generated from the surprisal-based account of pragmatic prediction:

(1) At the shape term, descriptions will be read more slowly in the unbiased conditions compared to the biased conditions. This will be the case for descriptions containing *einige* only if comprehenders interpret the quantifier as a pragmatic cue;

(2) At the shape term, descriptions containing *einige* will be read equally fast in both conditions, if comprehenders interpret *einige* as a semantic cue.

A visual representation of the predictions is provided in Figure 2.5.

2.4.3   Results

The data and analysis scripts are available for inspection under https://zenodo.org/record/5156186#.Y1kywYTMKUk.
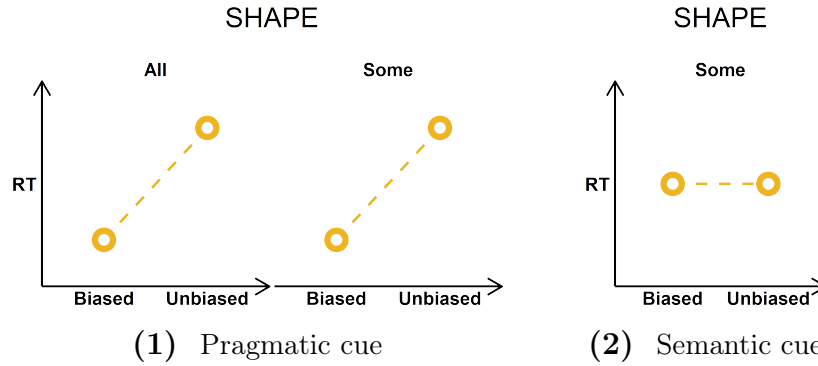
Figure 2.5 Study predictions.

Data cleaning

Participants' reading times (RTs) as well as their sentence ratings were recorded. Two participants were excluded from the original sample due to faulty data recording which resulted in incomplete data sets, meaning that prior to data cleaning the sample consisted of 56 participants. For each test sentence, RTs were measured across all words, for a total of 8 to 10 measurement regions per sentence depending on the sentence type (descriptions referring to the position of the shapes contained two additional regions compared to the other two types of description). The data was cleaned according to two criteria: first, for any given trial, if the total RT differed by 2.5 positive or negative standard deviations from the mean total reading time for the respective condition, then the trial was excluded from any subsequent analysis; then, for any given participant, if their number of excluded trials was superior to 30% of the total number of trials, then the participant was excluded from any subsequent analysis. While no participant was excluded on the basis of these criteria, 776 individual trials were excluded from any subsequent analysis. Moreover, inspection of the trial data inputted to the experimental program showed that in 1/3 of the Einige (False) items were coded erroneously, such that the critical term which rendered the descriptions false was the property term, when it should have been, in all cases, the shape term. In practice, what this particular sentence configuration did was to shift the incongruent critical term one region downstream. I therefore excluded the faulty trials from any subsequent

analysis.

Confirmatory analyses

Recall that on each trial participants were asked to rate the accuracy of the description they read using a 7-point scale. In the case of the biased and unbiased conditions, all trials consisted of semantically congruent and thus, a priori, accurate image-description pairs. In the case of the false conditions, all trials consisted of semantically incongruent and thus, a priori, inaccurate image-description pairs. The distribution of participants' ratings can be seen in Figure 2.6 below:
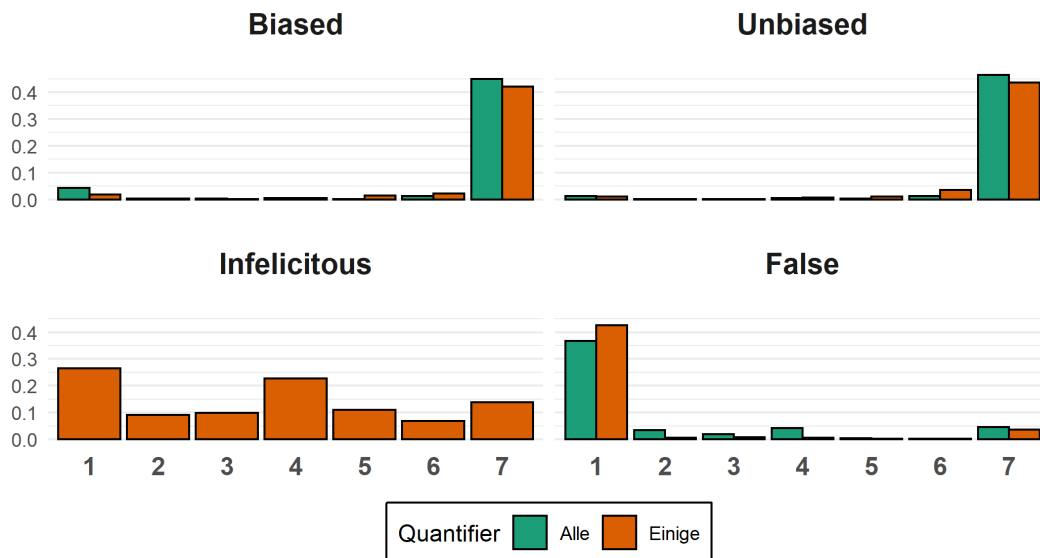
Figure 2.6 Sentence ratings. Each panel shows the proportion of ratings for a given critical condition. Ratings for sentences containing the quantifier *alle* are shown in green, while ratings for sentences containing the quantifier *einige* are shown in orange.

Visual inspection of the graph suggests that only rarely did participants rate the semantically congruent expressions as inaccurate descriptions of the pictures (rating 1, 2, and 3 on the 7-point scale). Similarly, only rarely did they rate the semantically incongruent expressions as accurate descriptions of the pictures (rating 5, 6, and 7 on the 7-point scale). Note that the ratings for the Einige (Infelict) condition are somewhat evenly distributed over

the whole scale, which means that there is wide variability in how participants interpreted descriptions containing the quantifier *einige*.

Figure 2.7 below shows the RT data. In order to amass quantitative evidence in favor of the reported results, I fitted Bayesian hierarchical models predicting RTs at the critical region, the SHAPE region, as a function of the experimental conditions, which themselves reflect different combinations of the quantifiers and the sentence bias. The models included, if possible, the maximal random effect structure justified by the design, which in the present case is random intercepts for items – the actual pictures seen by participants, which varied systematically according to the experimental condition – and random slopes and intercepts for participants. The models, fitted using the R package `brms` (Bürkner, 2017) and described in detail in the scripts available in the supporting material, had the following general form, shown in `brms` syntax:

$$\log(\text{RT}) \sim \text{condition} +$$
$$(1 + \text{condition} \mid \text{participant}) +$$
$$(1 \mid \text{item})$$

For each highlighted result, I report whether or not the respective statistical model provides strong evidence in favor of the empirically attested differences (or lack thereof). In a Bayesian statistical framework, one is interested in the joint posterior distribution of the parameters of the model, which indicates a plausible range of values for the parameters given the model and the data at hand. I report a 95% credibility interval (CI) and the posterior probability that the parameter of interest, β, is larger than zero ($P(\beta > 0)$). I speak of strong evidence for an effect when zero is not included in the CI and $P(\beta > 0)$ is close to either zero or one. Concretely, I'm interested in the difference between estimated values for cell means of conditions Einige (Biased) and Einige (Unbiased), as well as that between Alle (Biased) and Alle (Unbiased). Pragmatic surprisal theory predicts that, in the posterior distribution of the Bayesian regression model, the difference in cell means $\beta_{\text{einige}} =$ [estimates for cell

mean of Einige (Biased)] - [estimates for cell mean of Einige (Unbiased] should be credibly bigger than zero, so that $P(\beta_{\text{einige}} > 0)$ should be large, i.e., very close to 1; similarly for the conditions with the quantifier *all*.

Figure 2.7 shows the mean RTs at each sentence region up until the last critical term, across all seven experimental conditions. QUANT, SHAPE, and PROPER are the critical regions where task-relevant information was read, namely, the quantifier, the shape term, and the property term, respectively. Visual inspection of the graph suggests that there are differences between the critical conditions and the control conditions, which serve as a diagnostic of participants' online sensibility to the semantic congruency of the descriptions. At the SHAPE region, participants read the critical term faster in the critical conditions [orange and yellow] compared to the their respective controls [red and purple] (for the regression coefficients, see the respective table in the supporting material), as is to be expected.

As for the critical condition-quantifier pairs, visual inspection of the graph suggests that there are differences at the SHAPE region. Interestingly, however, the results go in the opposite direction of the predictions of pragmatic surprisal theory, contradicting my hypothesis: there is strong evidence that in the case of both *einige* and *alle* participants took longer reading the shape term in the biased condition compared to the unbiased condition. Indeed, the statistical models show that there is practically no reason to believe that the RTs in the unbiased conditions are larger than those in the biased conditions (β_{alle} = -0.06, 95% CI [-0.12, 0.00], $P(\beta_{\text{alle}} > 0)$ = 0.02; β_{einige} = -0.07, 95% CI [-0.13, 0.00], $P(\beta_{\text{einige}} > 0)$ = 0.02).

Summary

The results show that there were reliable differences in RTs at the SHAPE region, but these differences were in the opposite direction of what was predicted by pragmatic surprisal theory, which explains reading times as a monotone decreasing function of pragmatic expectability of the shape term in the context of the picture and the initial sentence segment. This theory predicts that both in the case of *einige* and *alle* RTs should be lower in the biased condition
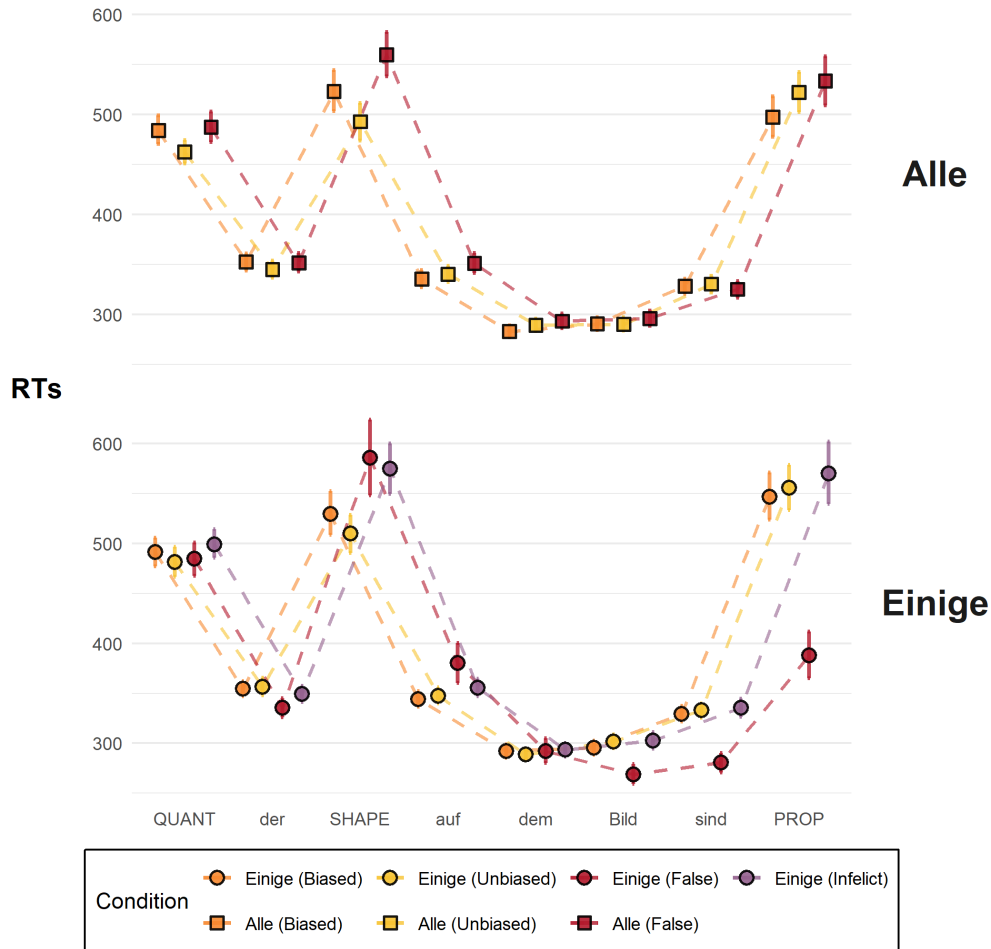
Figure 2.7 Mean reading times across all sentence regions up until the last critical term. The error bars represent 95% bootstrapped confidence intervals.

compared to the unbiased condition. However, in the case of both quantifiers, participants read the shape term more slowly in the biased condition compared to the unbiased condition. Table 2.2 summarizes the main results compared against the original predictions.

2.4.4   Discussion

All in all, the reported results warrant careful consideration. At face value, the observed pattern directly contradicts the predictions of pragmatic surprisal theory. But since PST consists of two components, the problem could lie with either component or both. Remem-

61

	Prediction		Result	
	Alle	*Einige*	*Alle*	*Einige*
SHAPE	Biased < Unbiased	Biased < Unbiased	Biased > Unbiased	Biased > Unbiased

Table 2.2 Summary of the results.

ber that PST assumes that (i) the reading times on a word are lower for more predictable words (the link function), and (ii) that probability of the next word is in turn influenced by contextual and pragmatic factors, in particular a preference for semantically true and pragmatically informative descriptions of the presented picture. Evidence against this conjunction of assumptions could be evidence against any one, or both, of these ideas.

Previous related work on pragmatic processing of scalar quantifiers in visually-anchored contexts (e.g. Augurzky et al., 2019; Spychalska et al., 2016) provides evidence for the idea that violations of pragmatic expectations, in the sense discussed here, do correspond with another assumed marker of next-word surprisal, namely the amplitude of an N400 component in ERPs (e.g. Kuperberg and Jaeger, 2016; Kutas and Federmeier, 2011). This suggests that context-induced pragmatic expectations, as conceived here, do seem to inform next-word expectations and may lead to surprisal-related processing difficulties.

There is also evidence that scalar implicature inferences (from *some* to *some but not all*) affect reading speed in self-paced reading studies. For example, Breheny et al., 2006 found increased reading times on a continuation with a phrase like *the rest* in contexts where a scalar implicature of a preceding occurrence of *some* was more expectable given a textual manipulation of expectation based on general world knowledge (similar results are presented by Bergen and Grodner, 2012). This suggests that self-paced reading times are, in principle, susceptible to pragmatic expectations of a sort.

Processing limitations

Given that there is some evidence for the ideas that (i) comprehenders do, at least sometimes, entertain pragmatic expectations of the kind of relevance here, induced by a visually presented context, and also that (ii) the self-paced reading method is susceptible to pragmatic factors, one possible explanation for the present results could be that the specific combination of the kind of pragmatic expectations (induced by a context picture), on the one hand, and self-paced reading, on the other hand, does, for some reason or other, not work. It could, for example, be that since self-paced reading is a less natural way of reading than on-screen reading in rapid visual serial presentation (as used, e.g., in comparable EEG studies), the burden on working memory of remembering a complex picture, forming (pragmatic expectations) and reading text in a self-paced manner is too onerous a task. In fact, in a recent study employing mouse-tracking, Darley et al. (2020) found that processing sentences against a visual context becomes more costly the higher the number of pragmatically-licensed sentence continuations, such that, in their study, higher numbers of possible continuations led to a decrease in task accuracy, an increase in the speed of responses, as well as higher degrees of attraction to foil responses in the measured mouse trajectories. The authors concluded that "[...] the main effects of the number of possible sentence completions observed here constitute evidence that episodic associations may be less conducive to the rapid and incremental incorporation of information and associated prediction-making that is made possible by a rich pragmatic context (perhaps specifically relying on long-term semantic associations or world knowledge)". Though in a very different setup, the task reported here also relies on episodic associations between visual stimuli and linguistic descriptions of those same stimuli, which is why it might be reasonable to raise similar issues in the present case as well.

But even if limited processing resources are an issue, this does not straightforwardly reconcile the present findings with pragmatic surprisal theory. The most natural effect of limited processing resources, on the assumption that the predictions of PST are basically

correct, would arguably be that the predicted differences in RTs would be deflated, perhaps to the extent that they are completely unattested in the data, i.e., one should expect lesser or no differences where PST predicts differences. However, the exact opposite pattern is attested in the data compared to what PST predicts. This is unexpected even if making room for limitations of processing resources due to the complexity of the task. In conclusion, blaming limited processing resources does not seem to vindicate PST in the light of the observed data.

Other pragmatic expectations

Pragmatic surprisal theory could be defended in the light of the obtained results by arguing that participants may have had different pragmatic expectations than the ones assumed throughout. This, however, it not a very convincing position given that the very same participants showed production behavior in the first part of the experiment which supports very directly the kind of pragmatic expectations assumed in above formulation of PST's predictions.

Task effects

Another class of potential alternative explanations to consider is task effects. There are at least two different kinds of task effects. For one, participants might adapt gradually to the statistical properties of the experimental environment, e.g., learning to associate a particular type of display with a particular kind of sentence and likely response. For another, there are task effects which do not require knowledge of the statistics of the experimental environments but constitute an approximation to a rational solution to the task as presented in the instructions. These two types of task-induced effects differ with respect to when during the course of the experiment they arise. While the former, frequency-driven effects are expected to emerge later during the experiment as participants acquire knowledge of the relevant statistics, the latter effects can, in principle, be expected already early on in the

experiment.
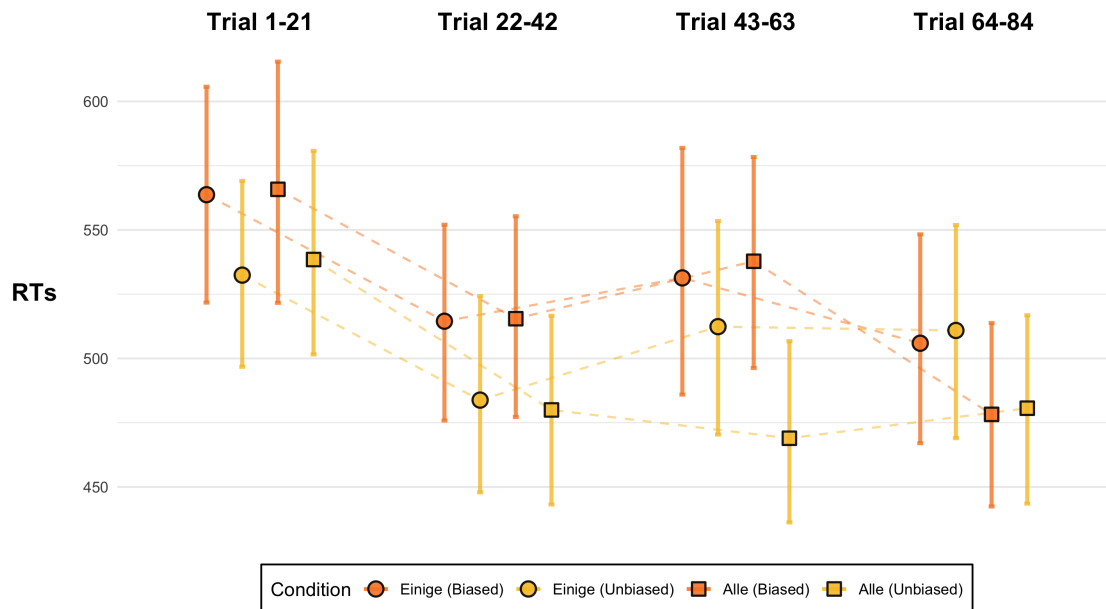
Figure 2.8 Reading times at the SHAPE region for the critical conditions for different blocks of the experiment. The dots show means, error bars are 95% bootstrapped confidence intervals.

Statistics of the experimental environment Pragmatic surprisal theory would not be discredited by the pattern observed in the aggregate data if this overall pattern could plausibly be explained as a task-effect based on the statistics of the experimental environment. To vindicate PST in this way, the predictions of PST should be borne out in the early parts of the experiment even if later parts of the experiment show emerging adaptive strategies leading away from the predicted behavior of PST. However, as seen in Figure 2.8, already during the first block of the experiment the main effect seen for the aggregate data – the opposite of what PST would predict – shows in the mean reading times (the supplementary material provides in-depth *post hoc* analysis of block effects). Indeed, at least numerically, the biased conditions are faster than the unbiased conditions across the whole experiment except the last block. There is no support for PST at the beginning of the experiment, nor in any other block. This suggests that vindicating PST by appeal to task effects that hinge on

participants adapting to the statistics of the experimental environment is not very plausible.

Strategic allocation of attentional resources   It remains to speculate about alternative *post hoc* explanations based on optimal solutions to the task without knowledge of the statistics of the environment. One conceivable alternative explanation for the faster reading of the unbiased conditions, compared to the biased conditions, revolves around *strategic allocation of attentional resources.* In this picture, what matters to self-paced reading speed is the immediate relevance of a chunk or word to the assessment of the pragmatic felicity of the unfolding descriptions. In other words, participants may be said to have read the shape terms more carefully, and thus more slowly, if the information at a given sentence region was relevant for assessing the pragmatic felicity of the description.

(a)   (b)
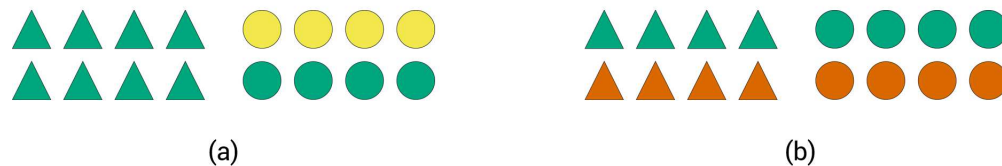
Figure 2.9 Biased (a) and unbiased (b) scenarios for descriptions containing the quantifier *einige.*

Consider, for instance, a scenario composed of a homogeneous triangle array and a heterogeneous circle array (Fig. 2.9 (a)). Having read *Einige der* (English: Some of the), even though a participant might be biased, as per the design, to expect *circles* next, she needs to know exactly whether *triangles* or *circles* are actually referred to – if *triangles*, the description is very likely to be either underspecified or downright false; if *circles*, then the load of determining pragmatic felicity is shifted to the subsequent critical region. Similarly, in the same scenario composed of a homogeneous triangle array and a heterogeneous circle array, having read *Alle der* (English: All of the), even though a participant might be biased, by design, to expect *triangles*, the shape term is key in determining whether the unfolding description is congruent or not: if *circles* are referred to, the description is rendered false at

that very sentence region; if, instead, *triangles* are referred to, then the load of determining the congruency of the description is shifted to the subsequent critical region.

Strategic allocation of attentional resources might explain the observed difference between unbiased conditions (read fast because they are irrelevant to the truth-judgement of the sentence) and biased conditions (potentially relevant information at the shape position to the truth-judgement of the sentence). This idea also explains why Einige (False) is read faster than Einige (Infelicitous). Notice that the context picture associated with Einige (False) is the same as that for Einige (Unbiased). Of course, to explain the increased reading times for conditions where the shape term makes the sentence (most likely) false, possibly by implicature, this alternative explanation must also stipulate a reading time increment for falsity.

2.5 Summary and conclusion

In this chapter I presented the results of a study designed to test the predictions of pragmatic surprisal theory. According to PST, visually-anchored contexts should induce pragmatic expectations about next-word continuations of sentences, and these pragmatic expectations should lead to increased processing efforts proportional to how unexpected incoming linguistic material is. PST is supported by previous research using EEG (e.g. Augurzky et al., 2019; Spychalska et al., 2016). As an alternative line of research also links next-word suprisal to reading times (e.g. Demberg and Keller, 2008; Monsalve et al., 2012; Smith and Levy, 2013), the study presented here aimed to test PST in the context of a self-paced reading study with visually-anchored contexts.

The observed results are in clear conflict with the predictions of PST. While limitations of processing resources might be relevant for this particular experimental design, it is not obvious how taking these into account would reconcile the empirical findings with PST. Based on *post hoc* inspection of the temporal development over the course of the experiment,

I argue that it is unlikely that the observed pattern, which I interpret as evidence against PST, is a task effect driven by the statistics of the experimental environment. I suggest an alternative *post hoc* explanation according to which reading times are a function of the strategic allocation of attentional resources, a process which is itself informed by context-induced pragmatic considerations but which does not rely on knowledge of the statistics of the experimental environment.

Ultimately, then, the present results seem to suggest that predictability in online pragmatic processing is linked not only to purely informativity-based prediction but also to other processing constraints, such as, in our case, one derived from a pressure to integrate crucial semantic information incrementally during processing. This is in line with constraint-based accounts of language comprehension which assume not only that multiple sources of information need to be integrated online during processing but perhaps more importantly that the weight of different constraints varies depending on the specific processing demands as well as on the larger discourse and communicative context (Degen and Tanenhaus, 2015). In fact, other studies on online pragmatic processing have also shown that pragmatic inferencing, including scalar inferencing cued by quantifiers, shows variable time courses and strong context-dependence (Bergen and Grodner, 2012; Huang and Snedeker, 2018; Urbach et al., 2015). What these different studies seem to agree on is that there is immense variability in how pragmatic interpretation works, from its contexts of occurrence, to its online signatures, and ultimately its underlying mechanistic processes. I conclude that more research is necessary to investigate how pragmatic expectations – as captured by the construal of PST formulated and tested here – combine or interact with resourceful task-dependent processing strategies.

REFERENCES

Augurzky, Petra, Michael Franke, and Rolf Ulrich (2019). "Gricean Expectations in Online Sentence Comprehension: An ERP Study on the Processing of Scalar Inferences". In: *Cognitive science* 43 8, e12776.

Bergen, Leon and Daniel J Grodner (2012). "Speaker knowledge influences the comprehension of pragmatic inferences." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 38.5, p. 1450.

Bicknell, K. and Hannah Rohde (2009). "Dynamic integration of pragmatic expectations and real-world event knowledge in syntactic ambiguity resolution". In:

Boudewyn, Megan A., Debra L. Long, and Tamara Y. Swaab (2015). "Graded expectations: Predictive processing and the adjustment of expectations during spoken language comprehension". In: *Cognitive, Affective, & Behavioral Neuroscience* 15, pp. 607–624.

Breheny, Richard, Napoleon Katsos, and John N. Williams (2006). "Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences". In: *Cognition* 100, pp. 434–463.

Bubić, Andreja, D. Yves von Cramon, and Ricarda I. Schubotz (2010). "Prediction, Cognition and the Brain". In: *Frontiers in Human Neuroscience* 4.

Bürkner, Paul-Christian (2017). "brms: An R package for Bayesian multilevel models using Stan". In: *Journal of statistical software* 80.1, pp. 1–28.

Clark, Andy (2013). "Whatever next? Predictive brains, situated agents, and the future of cognitive science." In: *The Behavioral and brain sciences* 36 3, pp. 181–204.

Darley, Emily, Christopher Kent, and Nina Kazanina (2020). "A 'no' with a trace of 'yes': A mouse-tracking study of negative sentence processing". In: *Cognition* 198.

Degen, Judith and Michael K Tanenhaus (2016). "Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study". In: *Cognitive science* 40.1, pp. 172–201.

— (2015). "Processing scalar implicature: A constraint-based approach". In: *Cognitive science* 39.4, pp. 667–710.

Demberg, Vera and Frank Keller (2008). "Data from eye-tracking corpora as evidence for theories of syntactic processing complexity". In: *Cognition* 109, pp. 193–210.

Gotzner, Nicole and Jacopo Romoli (2022). "Meaning and Alternatives". In: *Annual Review of Linguistics* 8.1, pp. 213–234.

Haase, Viviana, Maria Spychalska, and Markus Werning (2019). "Investigating the Comprehension of Negated Sentences Employing World Knowledge: An Event-Related Potential Study". In: *Frontiers in Psychology* 10.

Horn, Laurence R. (1989). "A Natural History of Negation". In:

Huang, Yi Ting and Jesse Snedeker (2018). "Some inferences still take time: Prosody, predictability, and the speed of scalar implicatures". In: *Cognitive psychology* 102, pp. 105–126.

Kamide, Yuki, Gerry TM Altmann, and Sarah L Haywood (2003). "The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements". In: *Journal of Memory and language* 49.1, pp. 133–156.

Kuperberg, Gina R and T Florian Jaeger (2016). "What do we mean by prediction in language comprehension?" In: *Language, cognition and neuroscience* 31.1, pp. 32–59.

Kutas, Marta and Kara D. Federmeier (2011). "Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP)." In: *Annual review of psychology* 62, pp. 621–47.

Levy, Roger (2008). "Expectation-based syntactic comprehension". In: *Cognition* 106.3, pp. 1126–1177.

Litwin, Piotr and Marcin Miłkowski (2020). "Unification by fiat: arrested development of predictive processing". In: *Cognitive Science* 44.7, e12867.

Mani, Nivedita and Falk Huettig (2012). "Prediction during language processing is a piece of cake–but only for skilled producers." In: *Journal of experimental psychology. Human perception and performance* 38 4, pp. 843–7.

Monsalve, Irene Fernandez, S. Frank, and Gabriella Vigliocco (2012). "Lexical surprisal as a general predictor of reading time". In: *EACL*.

Nieuwland, Mante S. (2016). "Quantification, Prediction, and the Online Impact of Sentence Truth-Value: Evidence From Event-Related Potentials". In: *Journal of Experimental Psychology. Learning, Memory, and Cognition* 42, pp. 316–334.

Nieuwland, Mante S., Yana Arkhipova, and Pablo Rodríguez-Gómez (2020). "Anticipating words during spoken discourse comprehension: A large-scale, pre-registered replication study using brain potentials". In: *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior* 133, pp. 1–36.

Nieuwland, Mante S., Tali Ditman, and Gina R. Kuperberg (2010). "On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities". In: *Journal of memory and language* 63.3, pp. 324–346.

Ridderinkhof, K Richard (2017). "Emotion in action: A predictive processing perspective and theoretical synthesis". In: *Emotion Review* 9.4, pp. 319–325.

Scappini, Maria et al. (2015). "On the non-incremental processing of negation: A pragmatically licensed sentence-picture verification study with Italian (dyslexic) adults". In:

Scholman, Merel C. J., Hannah Rohde, and Vera Demberg (2016). ""On the one hand" as a Cue to Anticipate Upcoming Discourse Structure". In:

Smith, Nathaniel J. and Roger Levy (2013). "The effect of word predictability on reading time is logarithmic". In: *Cognition* 128, pp. 302–319.

Spychalska, Maria, Jarmo Kontinen, and Markus Werning (2016). "Investigating scalar implicatures in a truth-value judgement task: evidence from event-related brain potentials". In: *Language, Cognition and Neuroscience* 31, pp. 817–840.

Urbach, Thomas P, Katherine A DeLong, and Marta Kutas (2015). "Quantifiers are incrementally interpreted in context, more than less". In: *Journal of memory and language* 83, pp. 79–96.

Winkler, István, Susan L Denham, and Israel Nelken (2009). "Modeling the auditory scene: predictive regularity representations and perceptual objects". In: *Trends in cognitive sciences* 13.12, pp. 532–540.

Chapter Three

Interpreting negated polar questions with epistemic biases and tracking beliefs in online discourse processing

3.1 Introduction

In Chapter 2, I presented a study focused on a well-studied class of pragmatic inferencing known as conversational implicatures, more specifically scalar implicatures. From a theoretical perspective, the derivation of inferences on the basis of scalars rests on the assumption that language users reason about alternatives, in this case lexical alternatives. Crucially, the relationship between scalar alternatives – like the quantifiers *some* and *all*, which were the focus of the study in Chapter 2 – might be pragmatically relevant in context as different expressions are more or less informative depending on their semantic status relative to one another. Pragmatic inferencing, however, need not be grounded in reasoning about lexical alternatives, such that other types of linguistic alternatives might provide interpreters with pragmatic cues to interpretation in context. In the present chapter, I tackle one such case of inferencing from non-scalar alternatives, more specifically inferencing from the (morpho)syntactic forms of questions and the epistemic biases that might be associated with

them. I start by reviewing related work on inferencing from information structure before discussing the issue of so-called questions with biases and introducing the empirical study which makes up this chapter.

3.1.1 Inferencing from information structure

Language users make use of a variety of devices to package information linguistically depending on their specific communicative needs. Indeed, depending on the particular language(s) one is using to communicate, devices can range from lexical and grammatical markers to specialized particles (Kučerová and Neeleman, 2012; Song, 2017) as well as various types of syntactic and prosodic modulations of the linguistic signal itself (Mithun, 2018; Song, 2017). These different devices are used – usually in combination with one another as well as with a variety of non-linguistic strategies – to structure information in an utterance in such a way that particular relevant meaning alternatives are highlighted, or such that particular pieces of information are signaled as (not) being part of the common ground.

These phenomena, particularly the ones operating at the level of (morpho)syntax and prosody, have been extensively studied in the domain of information structure (Krifka and Musan, 2012; see Féry and Ishihara (2016) for a comprehensive overview). Indeed, information structural notions such as *focus* – highlighting relevant (meaning) alternatives – and *givenness* – signaling what information is (not) part of the common ground – have been put forward as putative mechanisms underyling common cross-linguistic strategies for organizing linguistic utterances[1], their usage being guided by pragmatic considerations relating to how the content of a particular utterance fits within the unfolding discourse.

From a processing perspective, information structural cues have been shown to affect online language interpretation in context. On the one hand, syntactic cues have been shown

---

[1]Though see Ozerov (2021, 2018) for a recent proposal as for why these notions might need to be recontextualized in light of how and when information management strategies are actually employed in naturalistic language use.

to affect comprehenders' expectations with regard to possible discourse continuations, for instance when word order is used to predict upcoming discourse referents (e.g. Kaiser and Trueswell, 2004; Yano and Koizumi, 2018). On the other hand, prosodic cues are also known to be used to anticipate discourse, such as when pitch accent guides comprehenders towards not-yet-mentioned referents (e.g. Kurumada et al., 2014; Roettger and Franke, 2019). Common to these different instances of (predictive) inferencing is the fact that the relevant cues to interpretation are derived not primarily from the usage of particular lexical items or markers but rather from modulations in the form of an utterance which ultimately convey pragmatically relevant information. As discussed in Chapter 1, in a Gricean framework of pragmatic interpretation, inferences drawn on the basis of considerations related to the manner with which something is said are considered to be conversational implicatures, more specifically manner implicatures. Interpretation in such cases is assumed to be linked to more or less fine-grained sensitivities to do with what particular utterance formulations might signal in context.

What previous work on information structure suggests, then, is that interpretation can be guided by pragmatic considerations to do with the use of non-canonical utterance forms in context. However, while shedding light onto relevant aspects of pragmatic language interpretation, most research in the tradition of information structure has focused on the analysis of reference, more specifically the resolution of reference in declarative sentences, whose prototypical discourse function is to assert content. Yet assertions only make up one portion of the discursive landscape of language use: another essential aspect of pragmatic interpretation relates to how language users make sense of questions.

Questions have received a considerable deal of attention in research concerned with meaning in language, such as in semantics (e.g. Groenendijk and Stokhof, 1984; Karttunen, 1977), conversational analysis (e.g. G. Raymond, 2003; Stenström, 1984), and philosophy (e.g. Searle, 1969). Traditionally, the focus of research on questions has been on their canonical use, that is, how they function as requests for information or actions. A growing body of

research is now focused on so-called non-canonical questions (Dayal, 2016), which may differ from canonical ones both in terms of their form and function. From a pragmatic perspective, non-canonical questions are particularly interesting exactly as they do not conform to the standard definitions of what a question is, either in terms of form or function. Indeed, while questions are canonically associated with interrogatives – such as, in English, "Did you buy the present?" – they can also appear in declarative form – such as in indirect requests for information like "I was wondering if you bought the present". Questions can also serve functions which are not directly or primarily tied to information-seeking – such as in rhetorical questions, which are often claimed to be pragmatically equivalent to assertions of the opposite polarity (e.g. Han, 2002). For example, "Did you really need to buy such an expensive present?" seems to indicate that the questioner thinks that there was no need to buy an expensive present. Crucial for my current purposes, questions can also be non-canonical in the sense that, even though they do serve as requests for information, they are biased with regard to what the relevant answer(s) might be. In this chapter I focus on this latter dimension, namely how questions can carry biases and how these biases might impact incremental interpretation in context.

3.1.2 Questions with biases

While non-canonical questions in general have received attention across sub-disciplines of linguistics, from phonetics to syntax and semantics, questions with biases have been of particular interest to semantic and pragmatic research, both from a formal and a functional perspective (see De Ruiter (2012) as well as Romero (2020) for an overview and discussion). Much like other linguistic (pragmatic) phenomena related to meaning in discourse – like discourse markers, which are the focus of Chapters 4 and 5, questions with biases have been studied in relative isolation within different practice communities. And yet, despite varying epistemologies and little cross-talk between one another, these distinct communities seem to

converge at least on some theoretical assumptions about the intricate relationship between the form and function of questions in discourse.

In the interactional linguistic tradition, research has been focused on how questions can reflect particular stances which language users choose to take in communicative interaction, particularly in relation to the uncertaintity they associate with the states of affairs targeted in their questions (Heritage and C. Raymond, 2021; C. Raymond and Heritage, 2021). These (inter)subjective preferences in the design of questions – both in terms of their form and relative position in a particular interactional sequence – have also been shown to affect the preferences displayed by language users with respect to how they organize their *responses* to different sorts of questions (Heritage, 2012; Lee, 2015). Put together, these studies show that, at least in English, there are principled usage patterns whereby turns containing questions contingently affect subsequent turns-at-talk, including discourse moves containing new questions.

In the formal semantic tradition, on the other hand, the focus has been on the different sorts of (normative) biases a question can carry, and how those might, in turn, (co-)relate to different syntactic and prosodic realizations. I operationalize the research question addressed in the study described below primarily on the basis of previous work from this literature, although, as already noted, the core assumptions entertained here have also been discussed – albeit in different terms – in the interactional literature, pointing towards commonalities between otherwise radi(c)ally opposed research traditions.

One of the first and most far-reaching claims in the literature on questions with biases dates back to work by Robert Ladd in the 1980s (Ladd, 1981), where he discusses a 'puzzle' about negative polar questions which, in his view, is part of a larger puzzle on the scope of negation in questions that also includes tag questions, another type non-canonical question. Ladd speaks of a systematic ambiguity in negated polar questions, exemplified in (1) below (Ladd, 1981, p. 164):

(3)   **Context** Kathleen and Jeff have just come from Chicago on the Greyhound bus
        to visit Bob in Ithaca

        Bob: You guys must be starving. You want to go get something to eat?

        Kathleen: Yeah, isn't there a vegetarian restaurant around here – Moosewood, or
        something like that?

        Bob: Gee, you've heard of Moosewood all the way out in Chicago, huh? OK, let's
        go there.

As Ladd notes, "Kathleen uses the negative question *isn't there a vegetarian restaurant around here* to ask for confirmation of something she believes to be true", (Ladd, 1981, p. 164). He contrasts this situation with (2) below (Ladd, 1981, p. 164):

(4)   **Context** Bob is visiting Kathleen and Jeff in Chicago while attending CIS

        Bob: I'd like to take you guys out to dinner while I'm here — we'd have time to
        go somewhere around here before the evening session tonight, don't you think?

        Kathleen: I guess, but there's not really any place to go in Hyde Park.

        Bob: Oh, really, isn't there a vegetarian restaurant around here?

        Kathleen: No, about all we can get is hamburgers and souvlaki.

According to Ladd, "Bob uses the negated question here for a very different reason: he had previously assumed the truth of the proposition there is a vegetarian restaurant around here, but has now inferred from what Kathleen says that this proposition is actually false, and is using the negated question to check this new inference", (Ladd, 1981, p. 164). Indeed, in his view the first question – *isn't there a vegetarian restaurant around here* – is ambiguous between two possible meanings, such that "the negated question is being used to confirm something the speaker believes to be true", namely that there is a vegetarian restaurant around the area. In the second scenario, on the other hand, "the negated question is used to check on a new and unexpected inference", namely that there is no restaurant around the area.

Ladd's ambiguity, as the claim has come to be known, has been central to semantic work on polar questions, which have been addressed in a variety of theoretical accounts (e.g. Farkas and Roelofsen, 2017; Krifka, 2017; Van Rooy and Safarova, 2003). Crucial for my purposes, this claim has been recently put to the test in an experimental study by Domaneschi et al. (2017), one of the few empirical investigations on the matter. I follow Domaneschi et al.'s treatment of the theoretical literature, linking the relevant constructs of interest to their experimental results and in turn to my own study, which I introduce in the next section.

Both in the early work by Ladd (1981) as well as in more recent work by Romero and Han (2004), the usage of negated polar questions has been linked to the so-called *original speaker bias*. As an explanatory dimension of interest, original speaker bias relates to whether the producer of a question has any (prior) belief or expectation that the proposition targeted by their question is true, based on their epistemic state and/ or other relevant information which is part of the discourse situation. According to both Ladd (1981) and Romero and Han (2004), in English negated questions in which the negation marker *not* is contracted with the auxiliary (*n't*) – a so-called high negation – index a speaker bias, whereas negated questions in which the negation marker is realized separately from the auxiliary verb – a so-called low negation – index no particular bias from the side of the questioner. This is the first crucial element of the biased status of a polar question. The second element is known as *contextual evidence* and relates to the types of evidence possibly available in the discourse situation[2]. According to Büring and Gunlogson (2000), who first introduced the concept in this line of work, low negation questions are compatible with scenarios in which there is evidence against the relevant proposition, whereas high negation questions are incompatible with scenarios in which there is evidence against the relevant proposition, as exemplified in (3) below.

---

[2]See Romero and Han (2004) and the references therein for the original definitions of speaker bias and contextual evidence.

(5)   **Context** A enters S' windowless computer room wearing a dripping wet raincoat

(contextual evidence for p = it is raining)

S: What's the weather like out there? Is it raining?

S': #What's the weather like out there? Isn't it raining?

S": #What's the weather like out there? Is it not raining?

In this discourse situation, given the contextual piece of evidence – which suggests that it is raining, producing a straightforward non-negated polar question such as *Is it raining?* is a felicitous discourse move, while producing either form of a negated polar question – either with a high or low negation – is not, each for a different reason, as postulated by Büring and Gunlogson (2000). The relationship between the two bias elements – contextual evidence and original bias – has been treated differently by different authors, such that, depending on the particular account one takes as a starting point, their interaction can be said to be neither necessary nor sufficient for the felicitous usage of a biased yes/ no question. Empirically, however, the two factors have been found to interact to give rise to different usage profiles. Indeed, while, independently, each factor can be said to be either *biased for the proposition* (p), *neutral relative to the proposition* (neutral), or *biased against the proposition* (¬p), original and contextual bias can also be crossed with one another. The emerging picture can be visualized as a 3x3 design matrix, as in Table 3.1.

		Original bias		
		p	*neutral*	*¬p*
Contextual evidence	*p*			
	neutral			
	¬p			

Table 3.1 Theoretical design matrix of the bias in biased polar questions.

As per (3) above, it is clear that a negative question – regardless of its form – is rather

infelicitous in any of the cells in the first row of Table 3.1, as in a situation where there is a positive contextual bias for a proposition but the questioner does not have any prior expectations regarding the situation (top row, middle column). Having addressed the issue empirically, Domaneschi et al. investigated question production in the laboratory, both in English and German. While the original claim about negation in polar questions concerned question forms in English, it has since then been extended to other languages, including German (Romero and Han, 2004), Hugarian (Gärtner and Gyuris, 2017), and Japanese (Sudo, 2013).

Out of the languages claimed to show such biases, directly comparing English and German is particularly interesting as "though the syntax of the English and German polar question forms is largely parallel, their pragmatic use is partly misaligned" (Domaneschi et al., 2017, p. 10). Interestingly, while in English the high negation is said to be ambiguous between being neutral or biased against the relevant proposition, in German – where it's realized with the negation marker *nicht* – it is said to unambiguously signal neutral evidence for the proposition. So in a situation where someone is at a ticket office trying to buy a train ticket for the day after and they might think that there is a train in the early morning, an utterance like "Do you have any preference?" constitutes neutral evidence with regard to person's original belief, whereas "The only train available is at 11:00" is evidence against that belief. Now, in English, a question like *Isn't there a train in the early morning?* is said to be a felicitous discourse continuation to both "Do you have any preference?" and "The only train available is at 11:00", whereas its German equivalent *Gibt es nicht einen Zug früh am Morgen?* is said to be felicitous only as a response to a neutral antecedent like "Do you have any preference?".

In their production study, Domaneschi and colleagues asked participants to read questions aloud, selecting from different options presented to them on a list. Crucially, these questions were presented against discourse contexts in which original bias and contextual evidence were crossed with one another, allowing for a direct comparison between the two, as in the

design matrix presented in Table 3.1. While they investigated both negated and non-negated polar questions, here I focus only on the results concerning negated questions. What they found was that, both in English and in German, participants were more likely to produce questions with high negation – realized with *n't* and *nicht (ein)*, respectively – in contexts in which there was a particular belief or expectation about the proposition of interest, while there being contextual evidence against that very same proposition. That was also found to be the case when there was neutral contextual evidence for the proposition. As for questions with low negation – realized with *not* in English and *kein* in German, participants preferred producing them, again both in English and in German, in contexts in which the absence of a particular belief or expectation was combined with negative contextual evidence against the proposition. These results can be summarized by the empirical distribution shown in Table 3.2:

		Original bias		
		p	*neutral*	*¬p*
Contextual evidence	*p*			
	neutral	high negation		
	¬p	high negation	low negation	

Table 3.2 Attested pragmatic profile of negated polar questions, as per the data by Domaneschi et al. (2017).

These results confirm Ladd's original intuition that questions like *Isn't there a restaurant around here?* can be used in English to confirm a belief that something is true. By the same token, they also confirm his original intuition that the same questions can be used to disconfirm an assumption made on the basis of evidence against a positive belief. Domaneschi et al.'s results have since been replicated and extended upon in work on Italian (Di Maro et al., 2021), suggesting that signaling epistemic stance via negated questions might be a cross-linguistically common pragmatic strategy, at least for Indo-European languages. In-

terestingly, the Italian data also shows that there is a preference to produce high negation questions using the past tense, as opposed to the present, which raises questions about the exact pragmatic considerations underlying the empirically attested production preferences, and how those might relate to particular linguistic forms. All in all, the current empirical picture raises several interesting questions. First and foremost, from a theoretical perspective, one might ask what the implications of the attested production preferences might be for question interpretation, on the one hand, and for any underlying form representations, on the other. Can interpreters infer the relevant pragmatic biases from the forms of polar questions, or are the relevant epistemic meanings only weakly associated into them? If they are to some extent conventionally associated with particular polar questions, how do these meanings impact the processing of language and the interpretation of a speaker's commitment to the content of a question? Similarly to the study reported in Chapter 2, in the section below I report a study[3] which addresses the issue of whether pragmatic preferences evidenced in the production of negated polar questions affect how they are interpreted incrementally during online language comprehension. The reported study was pre-registered under https://osf.io/3zk6d and, unless otherwise noted, all analytic choices and procedures follow the pre-registered analysis plan.

3.2 Experiment 1 - Processing questions with epistemic biases in English

In the main task of this study, described below in section 3.2.2, I tested the extent to which comprehenders process negated polar questions drawing on pragmatic expectations about what the form of a negated question might encode about the epistemic status of the question producer. As explained in the previous section, my working assumptions are derived from

---

[3]This study has been conceived and carried out in collaboration with E Jamieson.

previous work in the literature suggesting that questions encode epistemic biases regarding what answer the questioner might expect to receive. In the present study, I focus my attention on negated polar questions in their canonical interrogative forms, my assumption being that, in English, questions with a so-called high negation form (*Didn't you buy the present?*) encode an expectation from the side of the questioner that whatever state of affairs they are questioning is indeed the case, given prior information they have regarding that particular state of affairs (or the probability thereof). In contrast, questions with a so-called low negation form (*Did you not buy the present?*) are said to signal no particular expectation of the questioner in relation to the state of affairs targeted by their question. Now, in order to test the hypothesis that the form of a negated question might signal the epistemic state of a question producer and in turn impact the incremental interpretation of a question, one needs to make sure to model situations in which the questions under consideration are embedded in discourse contexts which do in fact display the epistemic biases said to affect question interpretation. As discussed in the introduction, in naturalistic settings where language is used in communicative interaction the production of a particular question token is usually surrounded by rich interactional sequences as well as detailed discourse information, providing both questioner and answerer with layers of overt linguistic and non-linguistic behavior to anchor their cognitive processing to. In the contrived and over-simplified setting of a laboratory experiment, however, the situation tends to be radically different, which is why it becomes crucial to ensure that whatever dynamics are captured in an artificial laboratory task – despite being both qualitatively and quantitatively different from their naturalistic counterparts – still serve as meaningful proxies for the processes hypothesized to be at play in the wild. Given that my aim in this study was to try and systematically tease apart any potential effects of epistemic biases on online question interpretation, I first set out to ensure that the discourse contexts modeled in my processing task do indeed give rise to the epistemic expectations of interest. In other words, in the section below I report the results of a norming experiment used to create artificial discourse contexts which, crucially,

allows setting up theoretically meaningful expectations for the main task.

In this auxiliary norming task, I presented participants with excerpts from the written stimuli used in the main task described in section 3.2.2, namely the relevant contextual information which, by design, was meant to elicit particular beliefs about the modeled question producer. The aim behind this task was to collect independent norms of the experimental discourse contexts so as to know what the likelihood of expecting certain epistemic beliefs in those contexts might be. As such, the reported data indicates whether the items in the stimulus set elicit belief expectations which are in line with those postulated to have an effect on the main phenomenon of interest.

3.2.1 Norming task

Recall that the hypothesis tested in the main task relates to whether the syntactic form of a negated question in English, which is said to carry biases about the epistemic state of the question producer, impacts the incremental interpretation of such a question in context. Importantly, the key assumption behind this hypothesis is that the modulation of any given online processing patterns is contingent on the relevant questions being interpreted against contexts with a particular type of epistemic bias, namely contexts containing a prior belief about the proposition targeted by a question or contexts containing no particular belief about the proposition targeted by a question.

The norming task here reported was meant to tap into the naive understanding of English natives as to whether the specific discourse contexts constructed for the processing task do indeed give rise to biased interpretations about the presence or absence of a given questioner belief. Since no account of epistemic biases in questions currently available in the literature generates actual testable predictions about subjective interpretation of epistemic questioner beliefs, I stipulate, for my current purposes, my own predictions of what such interpretation patterns should look like. Contexts with a specific prior questioner belief, on the one hand,

should be naively interpreted as supporting whatever state of affairs is targeted by the question in consideration. For instance, in (4) below someone is adopting a new pet and the questioner arguably has a belief that that pet is a cat, given the information available in the discourse context.

(6) Our friend is adopting a new pet. You heard from her sister that it would be a cat.

Question: Isn't she/ Is she not adopting a cat?

Upon asking a naive interpreter whether or not they think the person in the scenario is adopting a cat, their response is likely 'yes', given that the context biases the positive state of affairs targeted by the question. Now compare this situation to a situation where there is no particular overt questioner belief, like in (5) below, where someone is telling someone else about a friend finishing a job and starting a new one.

(7) We are talking about a sold out concert that's coming up at our favorite venue.

Question: Didn't you/ Did you not buy a ticket?

Upon asking a naive interpreter whether or not they think the person in the scenario bought a ticket, their response is likely "Don't know" or "No", given the relative uncertainty about the actual state of affairs.

These different expectations relating to how the discourse context may impact naive interpreters' beliefs of the possible epistemic biases behind a negated question are key to the present study. In what follows I describe the task used to evaluate these expectations empirically.

Method

Participants

Forty self-reported native speakers of English were recruited using the crowdsourced platform Prolific. Participants were compensated monetarily for their participation in the experiment.

85

Data collection was conducted entirely online and all collected data was stored at servers from the Osnabrück University.

Materials and design

Participants were asked to read 20 fictional scenarios, each consisting of one or two short written statements relating to a common situation such as going to a restaurant or getting a new pet (e.g. "(4) Our friend is getting a new pet. You heard from her sister that it would be a cat."). Participants' task was to answer a question pertaining to each scenario, crucially, one which was a paraphrase of the relevant negated question investigated in the main experiment (e.g. "Do you think the person in this scenario is getting a cat?", paraphrasing "Isn't she/ Is she not getting a cat?"). There were 10 scenarios designed to contain a belief regarding the state of affairs targeted in the question and 10 designed to contain no explicit belief of any sort. The same three response options were available across all trials, namely "Yes", "No", or "Don't know". These options were counterbalanced across participants.

Procedure

Written instructions were provided prior to the actual task. Each trial, participants were presented with a scenario, a question pertaining to that scenario, as well as three response options, as described above. Participants were instructed to answer the question on the basis of the information available to them in the scenario. Upon clicking on one of the options and thus making a choice, participants were automatically directed to the next trial.

Results and discussion

The data and analysis scripts are available for inspection under https://osf.io/7a2jk/. Prior to the main experimental run (n=40), I piloted the norming task using the exact same procedure as in the main run in order to inspect the elicited response distributions and to diagnose any potential qualitative deviations from the interpretation patterns of interest.

Recall that the expectations were that *Prior belief* items should show a bias for "Yes" responses while *No belief* items should show a bias against "Yes" responses. Of course, these are qualitative predictions about the overall distribution of responses, as opposed to fine-grained quantitative predictions about the actual shape or range of the distributions. Because of that, there is no principled, theoretically-motivated way of demarcating what constitutes an actual bias, which is why I consider any ratio of "Yes" to "No"/ "Don't know" responses which is higher than 40/ 60, depending on the expected direction of the effect, as evidence for a bias in interpretation. Figure 3.1 shows the norms aggregated over all items and participants.
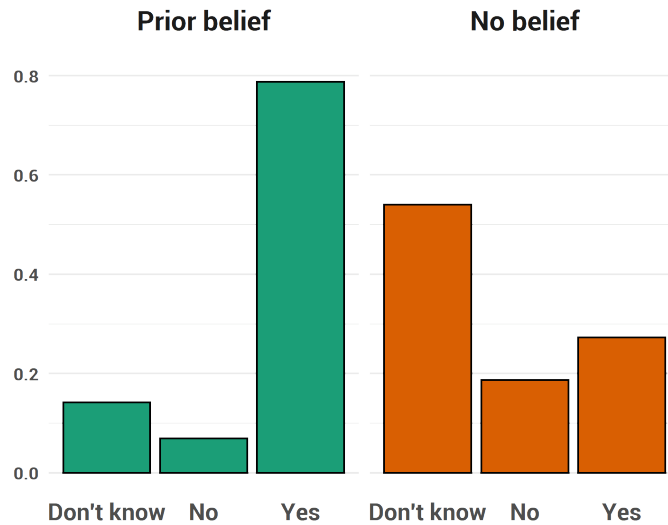
Figure 3.1 Item norms plotted by context type, *Prior belief* context shown in green, *No belief* context shown in orange. Each bar shows the proportion of choices of each response option, labeled individually below the respective bars.

As the graph shows, participants do interpret *Prior belief* items, on average, as reflecting specific beliefs about the discourse scenarios in those same items. By the same token, they do interpret *No belief* items, on average, as reflecting no particular prior beliefs about the discourse scenarios in those items. All in all, these results show that the normed items do reflect the general epistemic expectations which are of theoretical relevance to the account of question interpretation entertained in this study. Figure 3.2 shows an overview of the

individual norms, plotted in terms of each individual item in the item set, aggregated over all participants.
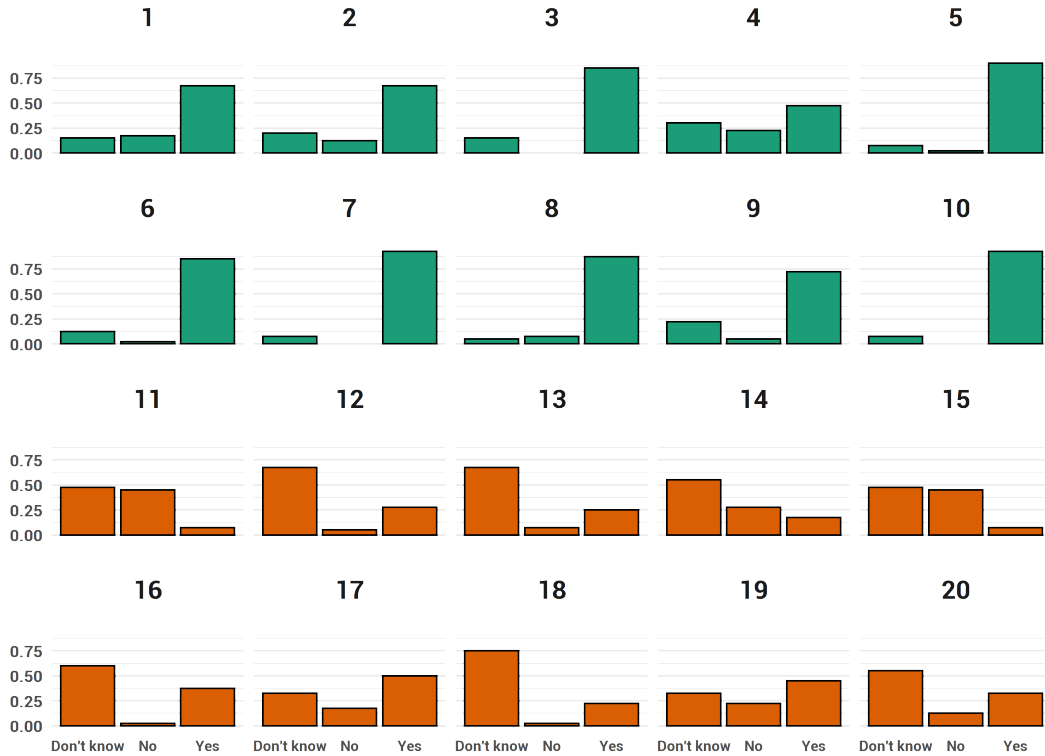
Figure 3.2 Item norms plotted by individual items. Each bar shows the proportion of choices of each response option, labeled individually below the respective bars.

The individual norms show a clear bias in the case of most items in the item set. There are, however, two items where there is no apparent bias for any particular epistemic expectation, namely items 4 and 19. For logistic reasons, I decided not to further modify any of these items. All in all, the norms show the natural degree of variation that one might expect in the sort of heterogeneous, naturalistic discourse text employed as stimuli in the study.

3.2.2 Processing task

In the main task of the present study, my goal was to look at how English natives interpret negated questions with epistemic biases incrementally during processing. The norming data presented in the section above shows that the discourse contexts built into the items

elicit the sort of epistemic expectations which are theoretically expected to affect question interpretation. I now turn to the details and setup of the experiment in which I investigated whether such contextually-relevant expectations modulate the online interpretation of negated questions.

Method

Participants

120 self-reported native speakers of English were recruited using the crowdsourced platform Prolific. Participants were compensated monetarily for their participation in the experiment. Data collection was conducted entirely online and all collected data was stored at servers from the Osnabrück University.

Materials and design

Participants were asked to read 35 fictional scenarios, each consisting of a short paragraph relating to a common situation such as going to a restaurant or getting a new pet as well as, crucially, a test sentence prefaced by another short sentence (see example item below). Participants' task was to read each scenario in its entirety, according to the procedure described below. There were 10 scenarios designed to contain a belief regarding the state of affairs targeted in the target sentence – which always consisted of a negated question – and 10 designed to contain no explicit belief of any sort. Every scenario also included some piece of information providing negative evidence regarding the state of affairs targeted in the target sentence. Any individual participant saw each scenario either with a high or a low negation question as a target. In addition to 20 the critical trials, there were 15 fillers containing non-negated questions as targets.

Example item (*Prior belief*, *high negation*)

Our friend is getting a new pet. You heard from her sister it would be a cat. However,

I tell you she's planning to take it for a lot of walks. You say:

Hold on. Isn't she getting a cat?

In the example item above, *You heard from her sister it would be a cat* is the belief, *However, I tell you she's planning to take it for a lot of walks* is the negative contextual evidence, and *Isn't she getting a cat?* is the target sentence.

Procedure

Written instructions were provided prior to the actual task, followed by three practice trials which mimicked the exact procedure of the test trials. On each trial, participants were presented with a short written scenario, as described above. The context paragraph was presented automatically at the center of the screen together with the written cue "Press the SPACE bar to reveal the words". Below the written cue there were underscores marking the location of the question and its preceding sentence, both of which were masked at the beginning of the trial. In order to reveal each word in the masked sentences, participants were instructed to press the space bar, as indicated by the cue on screen.

While the question was the actual linguistic stimulus of interest, the preceding sentence prevented the auxiliary verb in the question from being the very first region to be read on a trial. Crucially, participants read both sentences one word at a time, in a self-paced manner. Once they read the last region in the target sentence, pressing the space key would trigger a button labeled "Next", which had to be clicked in order to advance to the next trial. Once every third trial, instead of the "Next" button, participants were presented with a comprehension question relating to the scenario they read. They were asked to choose between two response options and received written feedback regarding their choice, which had the sole purpose of keeping participants engaged with the task of reading for comprehension.

Hypotheses

Recall that my working hypothesis was that the presence or absence of an epistemic bias would affect incremental question interpretation differentially depending on the syntactic form of the target questions. Importantly, this entails an interaction between epistemic bias and negation type, such that high negation questions are expected to be processed more easily in discourse contexts with a particular questioner belief compared to discourse contexts without any belief, whereas low negation questions are expected to be processed more easily in contexts without any questioner belief compared to contexts with a belief. Notice, however, that although the general prediction is that there should be differences between question types, and that these differences should emerge due to the processing of different negation forms, the account I'm entertaining is underspecified with regard to the precise locus and time-course of such a differential effect on online interpretation. In other words, the theory, as it stands, does not generate precise predictions as to when in processing such a difference should emerge. Should any difference emerge as a direct effect of processing the question form itself ($_{\text{high}}Didn't \ .../ \ _{\text{low}}Did \ you \ not \ ...$)? Or perhaps as an effect of processing, further downstream, the first semantic cue as to what the question is actually about ($_{\text{high}}Didn't \ you \ buy \ .../ \ _{\text{low}}Did \ you \ not \ buy \ ...$)? Complicating matters even more, the difference in form between a high negation and a low negation in English means that the two structures are not directly comparable when it comes to the potential locus of the expected effect. Whereas the high negation form consists of a contraction between the auxiliary verb used to mark a canonical question in English and the negation marker usually employed to reverse the polarity of the utterance (*Didn't*), in the low negation form these two elements are not only disjoint but there is also a third element, namely a pronoun, intervening between the two (*Did you not*), which has implications for how the negated question form is processed linearly as the linguistic signal unfolds.

Because of the issues just raised, at this stage in the investigation of negated questions

with biases it is an open, empirical question *whether* there is context-dependent facilitation during online processing and, if so, *when* exactly during incremental interpretation such facilitation occurs, leaving open considerable researcher degrees of freedom concerning where in the linguistic signal to situate the effect of interest. For the analysis here reported, I set the regions of interest as spanning over the question form and up until the main verb ($_\text{high}$*Didn't you buy*/ $_\text{low}$*Did you not buy*). This diverges from the regions of interest originally defined in the study pre-registration[4], which spanned only the question forms themselves ($_\text{high}$*Didn't*/ $_\text{low}$*Did you not*) along with their potential spillover regions, which in the current analysis are targeted as critical regions of interest. This revision in the analysis plan is motivated by two independent reasons. First, in line with the pre-registered plan, the main verb serves as the spillover region for the negation marker in the low negation, meaning that an effect caused by processing *not* might only be observed one region downstream from it, at the main verb. Then, as per the revised analysis, defined *after* originally inspecting the data, it seems reasonable to assume that the effect of processing the negated question form, both in the high and low negation, might only become apparent further downstream in the sentence when interpreting the first meaningful lexical cue as to what the actual content of the question is, or in other words, only as comprehenders process the first semantic cue as to what state of affairs the question is targeting (e.g. whether or not someone has brought an umbrella). Given this picture, in what follows I am primarily interested in looking at the processing signatures measured up until the main verb, the predictions being as stated at the beginning of this section.

Results

The data and analysis scripts are available for inspection under https://osf.io/7a2jk/.

---

[4]https://osf.io/3zk6d, see analysis section (C1) as well as the analysis script.

Data cleaning

Prior to any analysis, the data set was cleaned according to criteria defined in the pre-registered analysis plan, which include both participant-based and response-based exclusion criteria. I first excluded the data from any participant who self-reported being a monolingual speaker of a language other than English (n=2). I then further excluded any participant who did not reach an accuracy of 80% on the comprehension questions (n=4). Finally, I excluded any participant who deviated in more than 30% of their trials in more than 2.5 standard deviations from the grand mean for a given condition, as measured at any individual sentence region (n=1). Altogether, these exclusions resulted in a data set which included data from 113 participants.

In addition to excluding seven individual participants as per the criteria above, I excluded any individual data point which deviated in more than 2.5 standard deviations from the grand mean for a given condition (0.02% of the data). I also removed any data point lying outside of a pre-established range of interest set between 100 and 1000 ms (0.03% of the data). This reflects a meaningful range for the reading of individual words in self-paced reading (though see Jegerski and vanPatten (2013) on setting absolute cut-off points).

Confirmatory analyses

Before turning to the main results, which are based on aggregates over all experimental items and participants, I report and inspect the general reading time distributions. Crucially, recall that, by design, low negation items have one extra measurement region compared to high negation items, given the differences in syntactic form between both question types ($_{\text{high}}Didn't\ you\ buy$/ $_{\text{low}}Did\ you\ not\ buy$). As originally designed, the item set is also uneven
$\quad\quad\quad\quad\quad\quad\quad _1\quad\quad _2\quad _3\quad\quad\quad _1\quad _2\quad _3\quad _4$
with regard to the exact number of measurement regions per item, given the fact that the original regions of interest spanned only the question forms themselves, namely the contraction between the auxiliary and the negation marker in the high negation (*Didn't*)

and the disjoint auxiliary and negation marker in the low negation (*Did you not*). Despite not affecting the main analysis as per the original regions of interest, this imbalance in the total length of the sentences directly impacts the revised analysis which extends the regions of interest to the main verb, such that in shorter sentences the verb appears in the last sentence region (e.g. "$\underset{1}{Isn't}\ \underset{2}{Sarah}\ \underset{3}{\textbf{coming}}$?"), while in longer sentences it appears mid-sentence (e.g. "$\underset{1}{Won't}\ \underset{2}{she}\ \underset{3}{\textbf{have}}\ \underset{4}{a}\ \underset{5}{holiday}\ \underset{6}{in}\ \underset{7}{between}$?"). In order to match the items within each condition and avoid the conflation of sentence wrap-up effects with any potential effects of theoretical interest, I restrict my analysis of the original item set to items which have exactly five sentence regions in the high negation and exactly six sentence regions in the low negation, i.e., items which adhere to the sentence frame *auxiliary | pronoun | (not) | verb | verb spillover | noun*, as in "$\underset{1}{Hasn't}\ \underset{2}{she}\ \underset{3}{got}\ \underset{4}{a}\ \underset{5}{boyfriend}$?". For the remainder of this section, I report and analyze data from this reduced stimulus set, which consists of eight out of the original 20 critical items. In the next section of the present chapter, I report a second run of the experiment with redesigned items which adhere to the sentence frame above and thus have a balanced number of sentence regions within each condition. Figure 5.1 shows the empirical response distributions for each negation type at each sentence region.

As the figure shows, there is considerable overlap between the distributions at the sentence regions up until the main verb, which is to be expected as, even in the case of reliable effects, the usual mean word-level reading time differences in semantic and pragmatic processing are in magnitudes smaller than 50 ms. At the last sentence region, where the noun is presented, both in the case of high and low negation there is clear evidence for facilitation in *Prior belief* scenarios, which is where participants are expected to have strong contextual expectations about the noun they read, as opposed to *No belief* scenarios. Let's now turn to the grand means.

In order to amass quantitative evidence in favor of the results reported below, I fitted Bayesian hierarchical models predicting RTs at the different critical regions as a function of a nested effect of belief scenario within negation type. This allows directly testing the

Figure 3.3 Reading time distributions. Each column shows the response distributions for either the high (left column) or low negation items (right column), in both the *Prior belief* (green) and *No belief* (orange) scenarios, organized in panels according to the sentence region they originate from.

effect of condition for each negation type, as per the paragraph below. The models included the maximal random effect structure justified by the design, which in the present case is random intercepts for items – which varied systematically according to the belief scenario and negation type – and random slopes and intercepts for participants. These models, fitted using the R package `brms` (Bürkner, 2017) and described in detail in the scripts available in the supporting material, had the following general form, shown in `brms` syntax:

$$\log(\text{RT}) \sim \text{condition} \, / \, \text{negation} \, +$$
$$(1 + \text{condition} \, / \, \text{negation} \mid \text{participant}) \, +$$
$$(1 \mid \text{item})$$

For each highlighted result, I report whether or not the respective statistical model provides strong evidence in favor of the empirically attested differences (or lack thereof). In a Bayesian statistical framework, one is interested in the joint posterior distribution of the parameters of the model, which indicates a plausible range of values for the parameters given the model and the data at hand. I report a 95% credibility interval (CrI) and the posterior probability that the parameter of interest, β, is smaller than zero ($P(\beta < 0)$). One speaks of strong evidence for an effect when zero is not included in the CrI and $P(\beta < 0)$ is close to either zero or one. Concretely, I am interested in the difference between estimated values for cell means of the two conditions, *Prior belief* and *No belief*, for each negation type. The theory here entertained predicts that, in the posterior distribution of the Bayesian regression model, the difference in cell means β_{high} = [estimates for cell mean of *Prior belief*] - [estimates for cell mean of *No belief*] should be credibly smaller than zero, so that $P(\beta_{\text{high}} < 0)$ should be large, i.e., very close to 1; the opposite pattern is expected in the case of the low negation, that is, the difference in cell means β_{low} = [estimates for cell mean of *Prior belief*] - [estimates for cell mean of *No belief*] should be credibly bigger than zero, so that $P(\beta_{\text{low}} < 0)$ should be small, i.e., very close to 0.

Visual inspection of the results suggests that, in the high negation, there are no systematic differences between the relevant *Prior belief*/ *No belief* pairs at the two regions preceding the main verb (see regression coefficients in Table 5.1 below). At the main verb itself, the descriptive results suggest that *No belief* items are read more slowly than *Prior belief* items, in line with the prediction for the high negation. This result, however, finds only weak quantitative support in the respective statistical model (β_{high} = -0.16, 95% CrI [-0.38, 0.05], $P(\beta_{\text{high}} < 0) = 0.90$). In the low negation, visual inspection of the results suggests that, much like in the high negation, there are no systematic differences between the relevant *Prior belief*/ *No belief* pairs at the three regions preceding the main verb. At the main verb, the descriptive results suggest that *No belief* items are read more slowly than *Prior belief* items, contrary to what was predicted. This result finds strong quantitative support in the
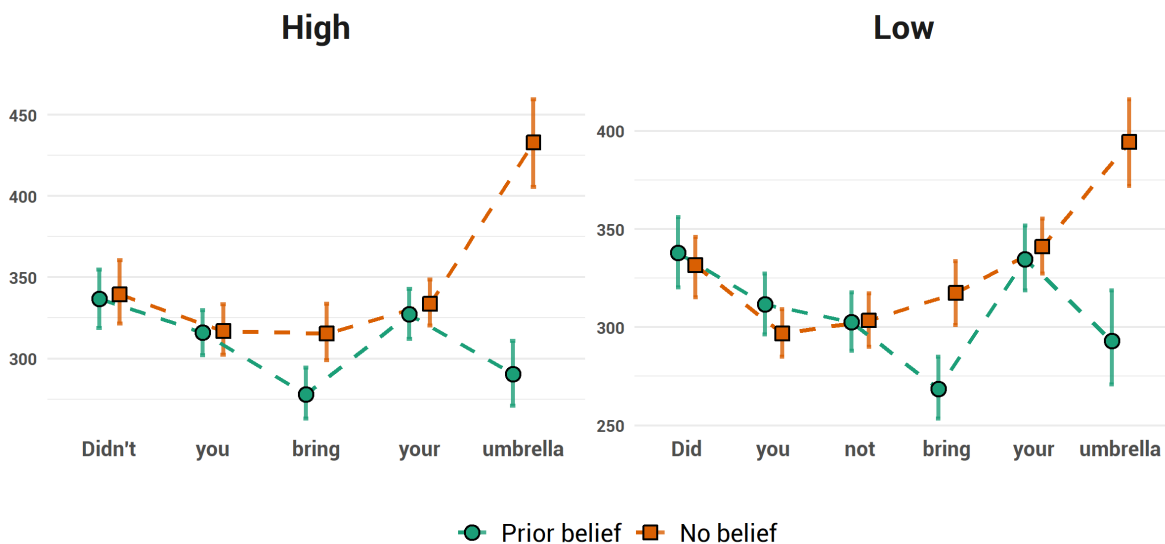
Figure 3.4 Mean reading times. Each column shows the grand mean for each each context type at each sentence region.

respective statistical model (β_{low} = -0.19, 95% CrI [-0.37, 0.00], $P(\beta_{\text{low}} < 0) = 0.95$).

Note that in the case of both high and low negation the last sentence region is read more slowly in *No belief* scenarios compared to *Prior belief* ones. This facilitation might be explained by the fact that, in the case of *Prior belief* items, the target questions are read against scenarios where, by design, the noun in the sentence is predictable given the discourse context. However, the effect might very well be an artifact of the imbalance in the materials described earlier. All in all, the current results do not support the original predictions, meaning that comprehenders do not seem to predict semantic material in a question based on pragmatic expectations to do with the epistemic state of the questioner. In the section below I report the results of a new experimental run where the imbalance in the original item set was fixed.

3.2.3 Processing task - Revised items

The original item set was fixed by revising any items which did not conform to the sentence frame *auxiliary | pronoun | (not) | verb | verb spillover | noun*, as in "*Hasn't she got a*

Region	Hypothesis	β	95%-CrI	P($\beta < 0$)
Auxiliary	[high] Prior belief < No belief	0.01	[-0.09, 0.08]	0.54
Pronoun	[high] Prior belief < No belief	-0.01	[-0.05, 0.04]	0.60
Verb	[high] Prior belief < No belief	-0.16	[-0.38, 0.05]	0.90
Verb spillover	[high] Prior belief < No belief	-0.04	[-0.22, 0.15]	0.65
Auxiliary	[low] Prior belief < No belief	0.00	[-0.08, 0.09]	0.48
Pronoun	[low] Prior belief < No belief	0.03	[-0.02, 0.08]	0.14
Negation marker	[low] Prior belief < No belief	-0.01	[-0.23, 0.21]	0.52
Verb	[low] Prior belief < No belief	-0.19	[-0.37, 0.00]	0.95

Table 3.3 Regression coefficients.

boyfriend?". Originally, only eight out of 20 items conformed to such a frame, which means that 12 items had to be revised. Changes included adding and/ or removing words from the critical sentences and in some cases slightly adapting the context sentence to better fit the revised questions. All other aspects of the study, including the procedure and the hypotheses, were the same as in the first experimental run. I again recruited 120 self-reported native speakers of English, following the same recruitment and data storage procedure as before.

Results

The data and analysis scripts are available for inspection under https://osf.io/7a2jk/.

Data cleaning

Prior to any analysis, the data set was cleaned according to the criteria defined in the pre-registered analysis plan, as in the first experimental run. I first excluded the data from any participant who self-reported being a monolingual speaker of a language other than English (n=5). I then further excluded any participant who did not reach an accuracy of at least 80% on the comprehension questions (n=2). Finally, I excluded any participant who deviated in

more than 30% of their trials in more than 2.5 standard deviations from the grand mean for a given condition, as measured at any individual sentence region (n=1). Altogether, these exclusions resulted in a data set which included data from 112 participants.

In addition to excluding eight individual participants as per the criteria above, I excluded any individual data point which deviated in more than 2.5 standard deviations from the grand mean for a given condition (0.01% of the data). I also removed any data point lying outside of a pre-established range of interest set between 100 and 1000 ms (0.02% of the data). This reflects a meaningful range for the reading of individual words in self-paced reading (though see Jegerski and vanPatten (2013) on setting absolute cut-off points).

Confirmatory analyses

Before turning to the main results, which are based on aggregates over all experimental items and participants, I report and inspect the general reading time distributions. Figure 3.5 shows the empirical response distributions for each negation type at each sentence region.

As the figure shows, there is considerable overlap between the distributions at the sentence regions up until the main verb, which is to be expected as, even in the case of reliable effects, the usual mean word-level reading time differences in semantic and pragmatic processing are in magnitudes smaller than 50 ms. Let's now turn to the grand means.

Visual inspection of the results suggests that, in the high negation, there are no systematic differences between the relevant *Prior belief* / *No belief* pairs at the two sentence regions preceding the main verb (see regression coefficients in Table 3.4 below). The same picture emerges at the main verb itself. In the low negation, much like in the high negation, there is no systematic difference between the relevant *Prior belief* / *No belief* pair at the auxiliary. At the pronoun, which follows the auxiliary and precedes the negation marker, the descriptive results suggest some facilitation for *No belief* items, as would be predicted for the low negation, however, this result finds no quantitative support in the statistical model (β_{low} = 0.01, 95% CrI [-0.02, 0.05], $P(\beta_{\text{low}} < 0) = 0.25$). Similarly, there seems to be some
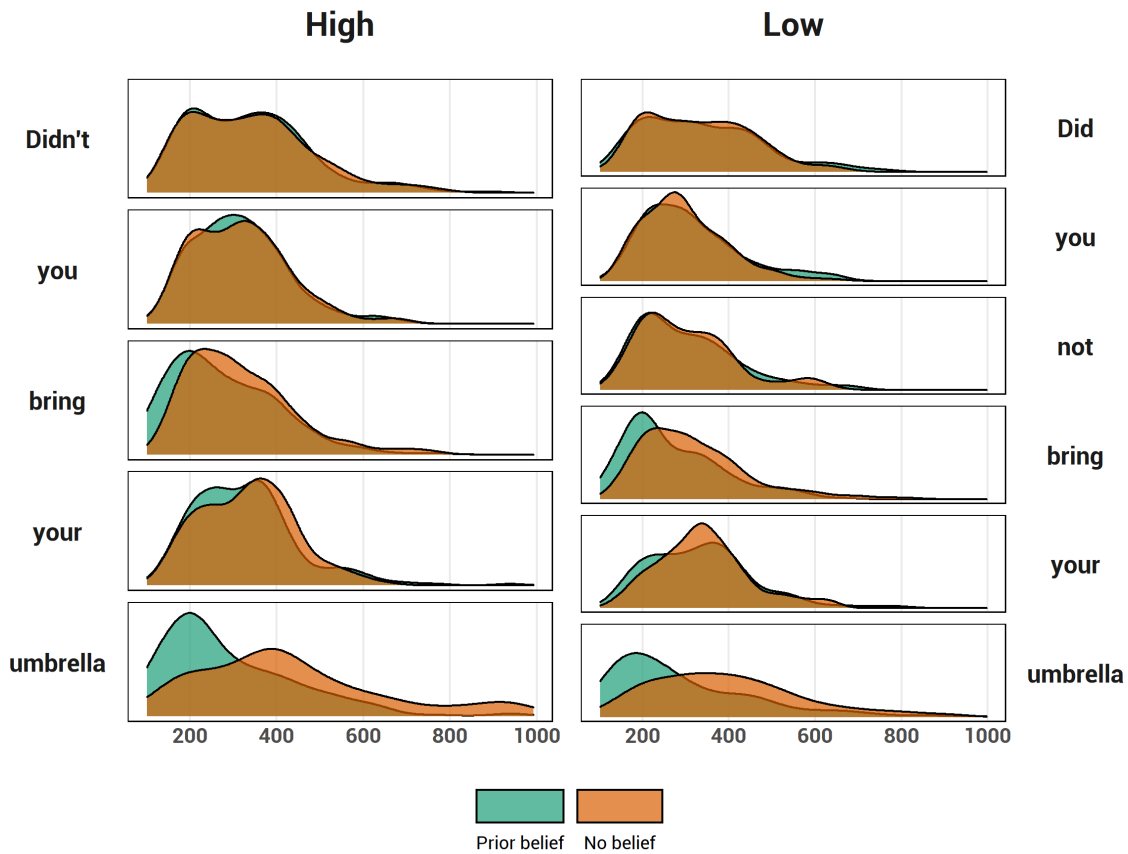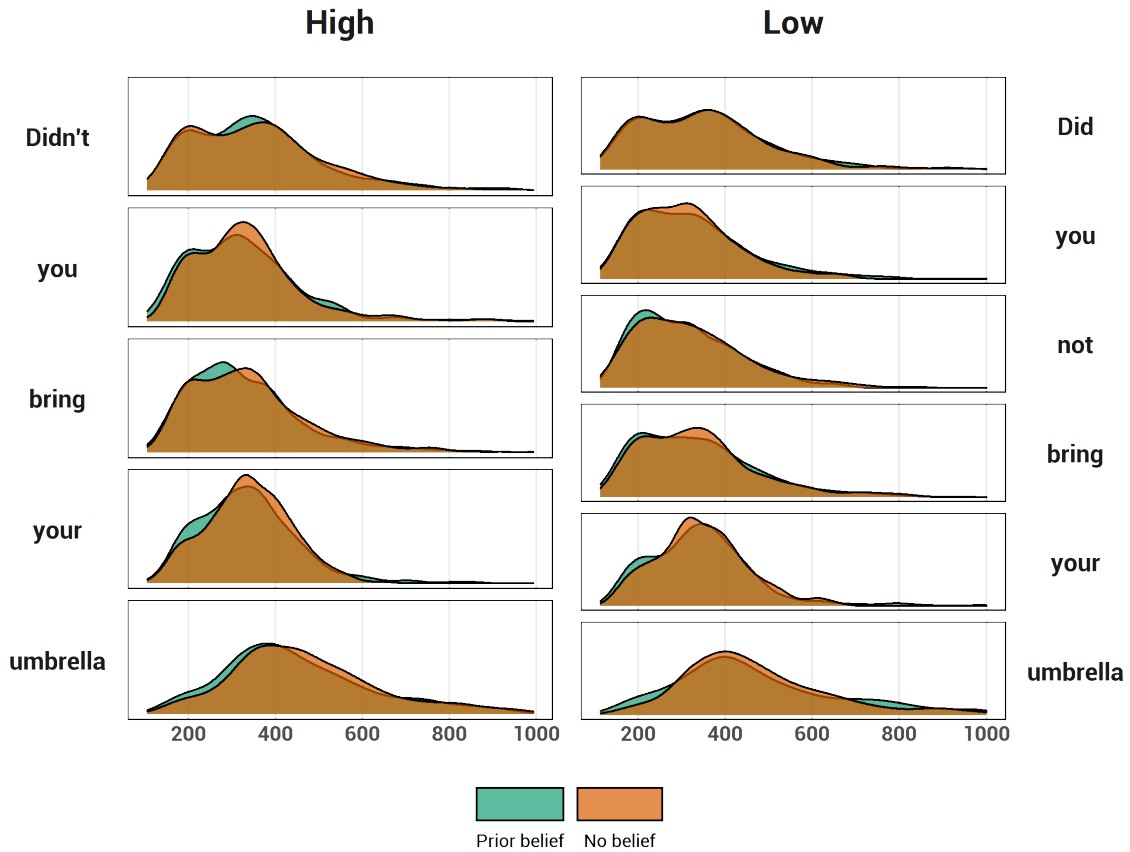
Figure 3.5 Reading time distributions. Each column shows the response distributions for either the high (left column) or low negation items (right column), in both the *Prior belief* (green) and *No belief* (orange) scenarios, organized in panels according to the sentence region they originate from.

facilitation for *Prior belief* items at the negation marker, which would run counter to the prediction for the low negation, however, this result finds no quantitative support in the respective model. Finally, at the main verb, there is no systematic difference between the relevant *Prior belief*/ *No belief* pair.

Note that in the case of high negation questions the last region in the sentence is read more slowly in *No belief* scenarios compared to *Prior belief* ones, a result which finds strong support in the respective statistical model (β_{high} = -0.07, 95% CrI [-0.12, -0.03], $P(\beta_{\text{high}} < 0) = 0.99$). This replicates, at least partially, the patterns seen in the first version of Experiment 1, though no systematic difference is attested in the case of low negation
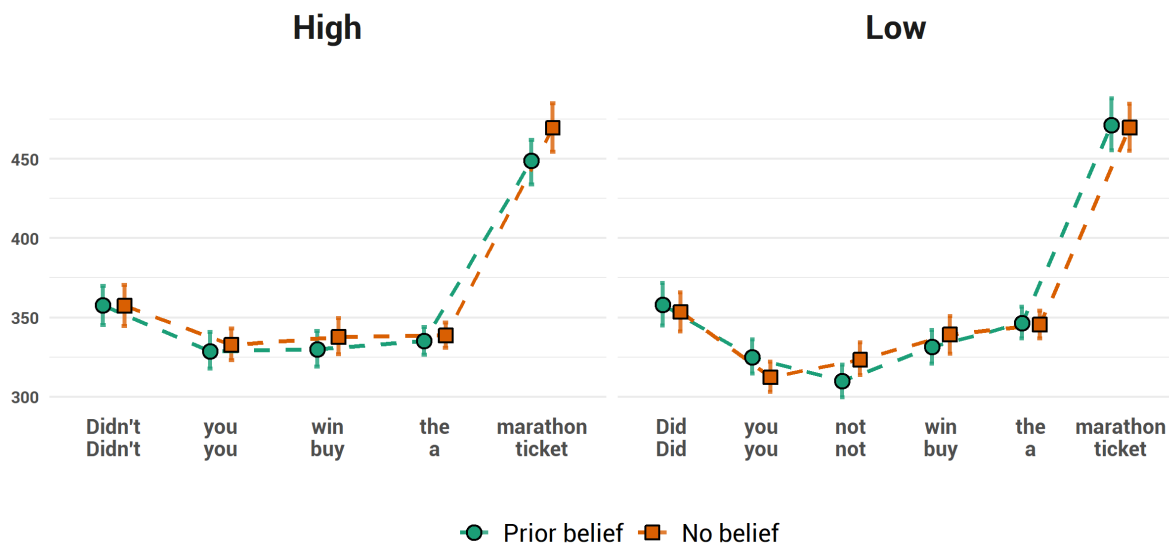
Figure 3.6 Mean reading times. Each column shows the grand mean for each each context type at each sentence region.

questions.

3.3 Discussion

In the present study, I set out to investigate the incremental interpretation of questions with biases, focusing on questions in English. Crucially, I addressed this issue using a task in which participants read negated polar questions against discourse contexts which contained different epistemic biases relating to the propositions targeted in the questions. Biases were operationalized as a combination of contextual evidence against the relevant proposition, on the one hand, and the presence or absence of a particular belief pertaining to the proposition, on the other. My predictions were that these biases would interact with the form of a question – more specifically, the syntactic form of the negated questions – to give rise to differential effects on the measured reading times.

More concretely, I expected reading times at the critical sentence regions – which spanned from the auxiliary verb (*Didn't/ Did*) to the main verb (Didn't/ Did you not *buy*) – to be

Region	Hypothesis	β	95%-CrI	P($\beta < 0$)
Auxiliary	[high] Prior belief < No belief	0.00	[-0.03; 0.04]	0.42
Pronoun	[high] Prior belief < No belief	-0.03	[-0.06; 0.00]	0.94
Main verb	[high] Prior belief < No belief	-0.04	[-0.08; 0.00]	0.93
Verb spillover	[high] Prior belief < No belief	-0.03	[-0.07; 0.00]	0.96
Auxiliary	[low] Prior belief < No belief	0.01	[-0.03; 0.05]	0.32
Pronoun	[low] Prior belief < No belief	0.01	[-0.02; 0.05]	0.25
Negation marker	[low] Prior belief < No belief	-0.02	[-0.06; 0.02]	0.79
Main verb	[low] Prior belief < No belief	-0.02	[-0.05; 0.01]	0.84

Table 3.4 Regression coefficients.

lower when a high negation question was read in a discourse context containing a particular belief, and similarly so when a low negation question was read in a discourse context containing no particular belief. The hypothesized link between the empirical reading signatures and the underlying processing mechanism states that a particular word is harder to process – and hence takes longer to read – the less predictable that word is in context, in my case predictability being, by design, assumed to be tied to explicit expectations as per the constraints described above.

The quantitative analyses reported in the previous sections do not support these predictions. Interestingly, despite not supporting the predictions with respect to the regions of interest, the results show evidence of facilitatory processing at the last sentence region in high negation questions read against contexts containing a questioner belief. I interpret this empirical observation as evidence that comprehenders might be able to track epistemic information when processing negated polar questions in English, despite not being sensitive to how that information might interact with different negation forms and the pragmatic biases arguably associated with them, at least not as evidenced in the current data.

3.4 Experiment 2 - Processing questions with epistemic biases in German

Experiment 2 follows the design and procedure of Experiment 1, focusing on the processing of biased questions in German. The reported study was pre-registered under https://osf.io/hk3pb and, unless otherwise noted, all analytic choices and procedures follow the pre-registered analysis plan. Just like in Experiment 1, I first report the results of the norming task.

3.4.1 Norming task

Method

Participants

Forty self-reported native speakers of German were recruited using the crowdsourcing platform Prolific. Participants were compensated monetarily for their participation in the experiment. Data collection was conducted entirely online and all collected data was stored at servers from the Osnabrück University.

Materials, design, and procedure

The design and procedure were the same as in the norming task used in Experiment 1. The materials were translated from English into German and modified accordingly so that they not only reflected the right epistemic expectations but also sounded pragmatically appropriate.

Results and discussion

The data and analysis scripts are available for inspection under https://osf.io/2ndf3/. As in Experiment 1, I ran three norming pilots, each with n=5, before running the full norming

task with n=40. Figure 3.7 shows the results of the full experimental run, showing the norms aggregated over all items and participants.
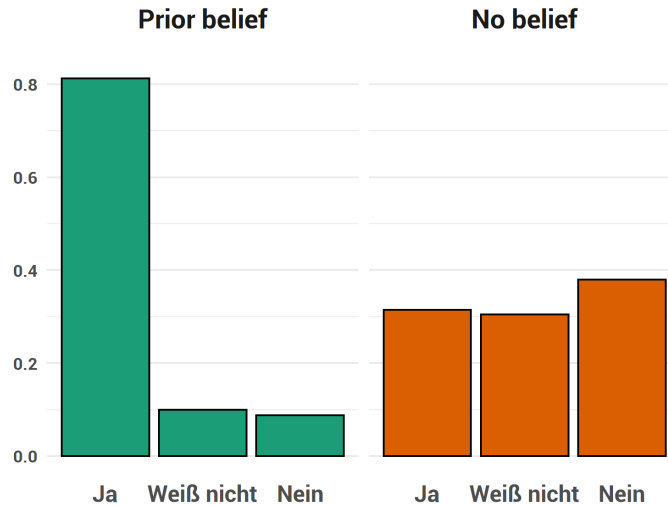
Figure 3.7 Item norms plotted by context type, *Prior belief* context shown in green, *No belief* context shown in orange. Each bar shows the proportion of choices of each response option, labeled individually below the respective bars.

As the graph shows, participants do interpret *Prior belief* items, on average, as reflecting specific beliefs about the discourse scenarios in those same items. By the same token, they do interpret *No belief* items, on average, as reflecting no particular prior beliefs about the discourse scenarios in those items. All in all, these results show that the normed items do reflect the general epistemic expectations which are of theoretical relevance to the account of question interpretation entertained in this study. Figure 3.8 shows an overview of the individual norms, plotted in terms of each individual item in the item set, aggregated over all participants.

The individual norms show a clear bias in the case of most items in the item set. There is, however, one item where there is no apparent bias for any particular epistemic expectation, namely item 4, as well as an item where the bias is in the opposite direction of what was expected, namely item 19. For logistic reasons, I decided not to further modify any of these items. All in all, the norms show the natural degree of variation that one might expect in
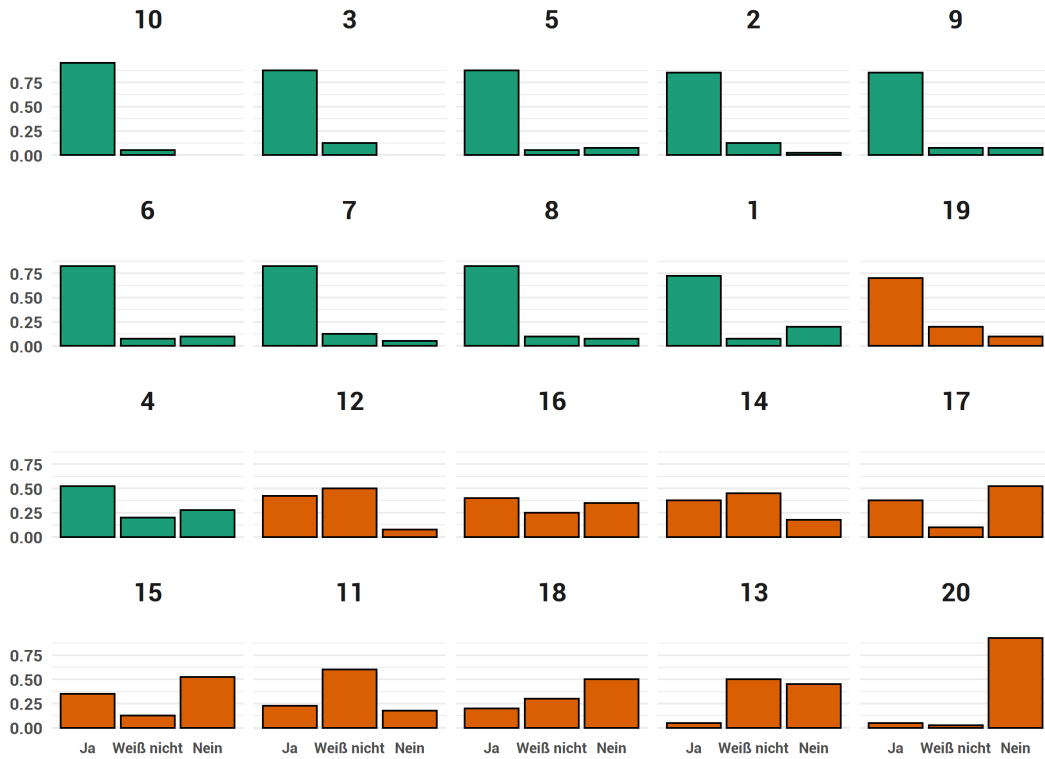
Figure 3.8 Item norms plotted by individual items. Each bar shows the proportion of choices of each response option, labeled individually below the respective bars.

the sort of heterogeneous, naturalistic discourse text employed as stimuli in the study.

3.4.2 Processing task

Method

Participants

80 native speakers of German were recruited among the cognitive science student population of the Osnabrück University. Participants were given course credit for their participation in the experiment. Data collection was conducted entirely online and all collected data was stored at servers from the Osnabrück University.

Materials, design, and procedure

The design and procedure were the same as in the processing task used in Experiment 1. The materials consisted of translated and modified versions of the English items used in Experiment 1, as explained above.

Hypotheses

My working hypothesis was the same as in Experiment 1, namely that the presence or absence of an epistemic bias would affect incremental question interpretation differentially depending on the syntactic form of the target questions. Importantly, this entails an interaction between epistemic bias and negation type, such that high negation questions are expected to be processed more easily in discourse contexts with a particular questioner belief compared to discourse contexts without any belief, whereas low negation questions are expected to be processed more easily in contexts without any questioner belief compared to contexts with a belief. Notice that, much like in English, the two question forms in German do not have the same number of words. Whereas the high negation form consists of a sentential negation where the negation marker is placed before an indefinite article (*Hast du nicht ein*), the low negation form consists of a morphologically negated indefinite article (*Hast du kein*).

Just like in the revised version of Experiment 1, the regions of interest span over the question form and up until the firs semantic cue in the sentence, which in German was a noun as opposed to the main verb ($_{\text{high}}$*Hast du nicht ein Ticket*/ $_{\text{low}}$*Hast du kein Ticket*). The predictions are as stated at the beginning of this section.

Results

The data and analysis scripts are available for inspection under https://osf.io/2ndf3/.

106

Data cleaning

Prior to any analysis, the data set was cleaned according to criteria defined in the pre-registered analysis plan, as in Experiment 1. I first excluded the data from any participant who did not reach an accuracy of 80% on the comprehension questions (n=2). Finally, I excluded any participant who deviated in more than 30% of their trials in more than 2.5 standard deviations from the grand mean for a given condition, as measured at any individual sentence region (n=1). Altogether, these exclusions resulted in a data set which included data from 77 participants.

In addition to excluding three individual participants as per the criteria above, I excluded any individual data point which deviated in more than 2.5 standard deviations from the grand mean for a given condition (0.02% of the data). I also removed any data point lying outside of a pre-established range of interest set between 100 and 1000 ms (0.01% of the data). This reflects a meaningful range for the reading of individual words in self-paced reading (though see Jegerski and vanPatten (2013) on setting absolute cut-off points).

Confirmatory analyses

Before turning to the main results, which are based on aggregates over all experimental items and participants, I report and inspect the general reading time distributions. Crucially, recall that, by design, high negation items have one extra measurement region compared to low negation items, given the differences in syntactic form between both question types ($_{\text{high}}Hast\ du\ nicht\ ein/\ _{\text{low}}Hast\ du\ kein$). Figure 3.9 shows the empirical response distributions for each negation type at each sentence region.

As the figure shows, there is considerable overlap between the distributions at all sentence regions, which is to be expected as, even in the case of reliable effects, the usual mean word-level reading time differences in semantic and pragmatic processing are in magnitudes smaller than 50 ms. Let's now turn to the grand means.
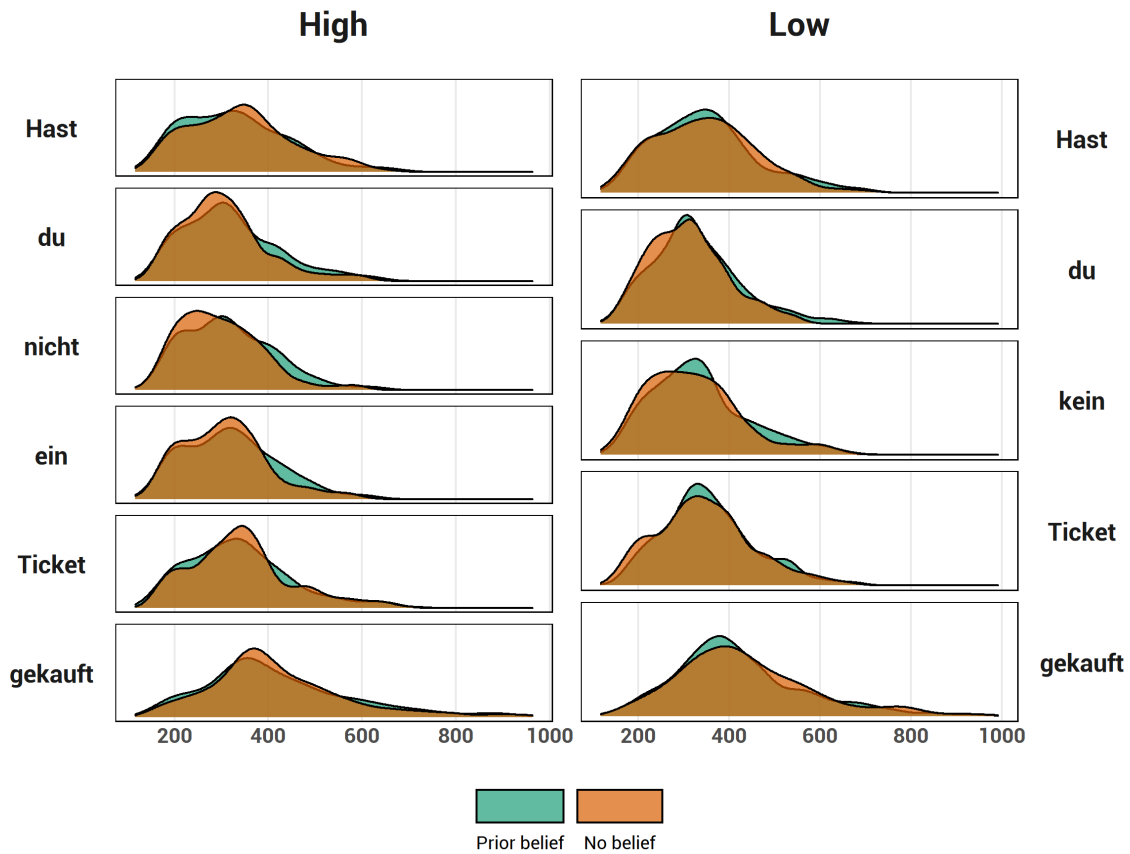
Figure 3.9 Reading time distributions. Each column shows the response distributions for either the high (left column) or low negation items (right column), in both the *Prior belief* (green) and *No belief* (orange) scenarios, organized in panels according to the sentence region they originate from.

In order to amass quantitative evidence in favor of the results reported below, I fitted Bayesian hierarchical models predicting RTs at the different critical regions as a function of a nested effect of belief scenario within negation type. This allows directly testing the effect of condition for each negation type, as per the paragraph below. The models included the maximal random effect structure justified by the design, which in the present case is random intercepts for items – which varied systematically according to the belief scenario and negation type – and random slopes and intercepts for participants. These models, fitted using the R package `brms` (Bürkner, 2017) and described in detail in the scripts available in the supporting material, had the following general form, shown in `brms` syntax:

$$\log(\text{RT}) \sim \text{condition} \ / \ \text{negation} \ +$$
$$(1 + \text{condition} \ / \ \text{negation} \ | \ \text{participant}) \ +$$
$$(1 \ | \ \text{item})$$

For each highlighted result, I report whether or not the respective statistical model provides strong evidence in favor of the empirically attested differences (or lack thereof). In a Bayesian statistical framework, one is interested in the joint posterior distribution of the parameters of the model, which indicates a plausible range of values for the parameters given the model and the data at hand. I report a 95% credibility interval (CrI) and the posterior probability that the parameter of interest, β, is smaller than zero ($P(\beta < 0)$). One speaks of strong evidence for an effect when zero is not included in the CrI and $P(\beta < 0)$ is close to either zero or one. Concretely, I am interested in the difference between estimated values for cell means of the two conditions, *Prior belief* and *No belief*, for each negation type. The theory here entertained predicts that, in the posterior distribution of the Bayesian regression model, the difference in cell means $\beta_{\text{high}} = $ [estimates for cell mean of *Prior belief*] - [estimates for cell mean of *No belief*] should be credibly smaller than zero, so that $P(\beta_{\text{high}} < 0)$ should be large, i.e., very close to 1; the opposite pattern is expected in the case of the low negation, that is, the difference in cell means $\beta_{\text{low}} = $ [estimates for cell mean of *Prior belief*] - [estimates for cell mean of *No belief*] should be credibly bigger than zero, so that $P(\beta_{\text{low}} < 0)$ should be small, i.e., very close to 0.

Visual inspection of the results suggests that, in the high negation, *Prior belief* items are read faster than *No belief* items at the auxiliary verb, in line with the prediction for the high negation. This result, however, finds only weak quantitative support in the respective statistical model ($\beta_{\text{high}} = $ -0.03, 95% CrI [-0.08, 0.01], $P(\beta_{\text{high}} < 0) = 0.87$). At the three regions following the auxiliary and preceding the noun, the results show that *No belief* items are read faster than *Prior belief* items, contrary to the prediction for the high negation. All these results find strong quantitative support in the respective statistical models (see
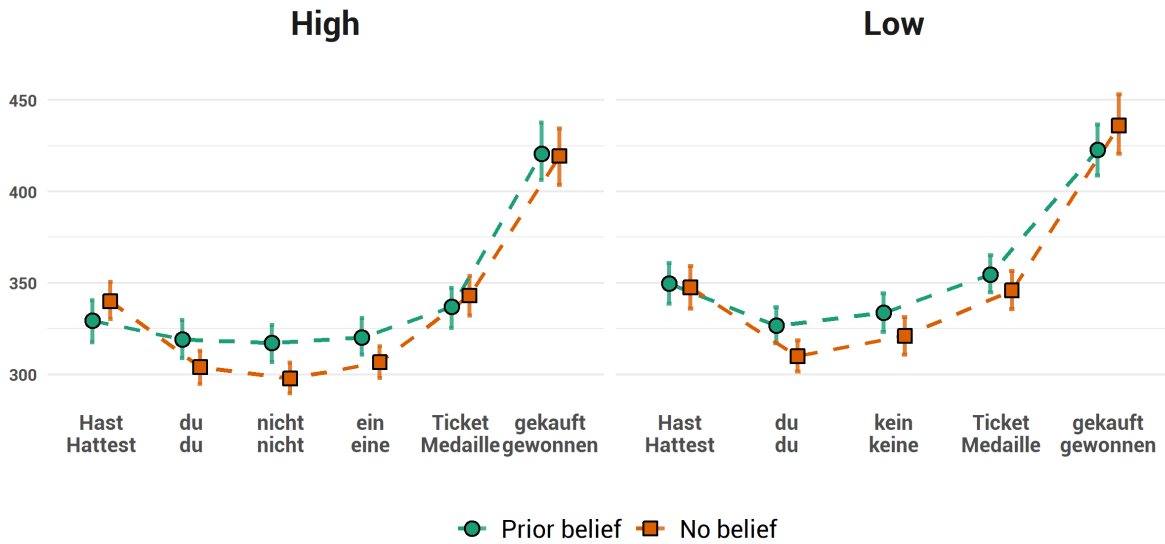
Figure 3.10 Mean reading times. Each column shows the grand mean for each each context type at each sentence region.

regression coefficients in Table 3.5 below). At the noun, the descriptive result suggests that there is no systematic difference between the *Prior belief* / *No belief* pair. In the low negation, visual inspection of the results suggests that there is no systematic difference between the relevant *Prior belief* / *No belief* pair at the auxiliary. At the two regions following the auxiliary and preceding the noun, the results show that *No belief* items are read faster than *Prior belief* items, in line with the prediction for the low negation (see regression coefficients in Table 3.5 below). Notice, however, that the result at the article finds weaker quantitative support compared to the result at the pronoun (see regression coefficients in Table 3.5 below). At the noun, while the descriptive result might suggest that *No belief* items are read faster than *Prior belief* items, in line with what was predicted, the result finds no quantitative support in the respective statistical model ($\beta_{\text{high}} = 0.01$, 95% CrI [-0.03, 0.06], $P(\beta_{\text{high}} < 0) = 0.32$).

While I did not have any predictions for the reading patterns at the main verb, the last region in the sentence, given the results of Experiment 1, one might expect that in German high negation questions the verb would also be read faster in *Prior belief* scenarios

compared to *No belief* ones. This is, however, not the case, even though in the low negation the descriptive result seem to suggest an effect (which is not supported by the statistical model).

Region	Hypothesis	β	95%-CrI	P($\beta < 0$)
Auxiliary	[high] Prior belief < No belief	-0.03	[-0.08, 0.01]	0.87
Pronoun	[high] Prior belief < No belief	0.04	[0.00, 0.08]	0.03
Negation marker	[high] Prior belief < No belief	0.05	[0.01, 0.09]	0.03
Article	[high] Prior belief < No belief	0.05	[0.01, 0.08]	0.02
Noun	[high] Prior belief < No belief	-0.04	[-0.23, 0.15]	0.66
Auxiliary	[low] Prior belief < No belief	0.01	[-0.03, 0.06]	0.32
Pronoun	[low] Prior belief < No belief	0.05	[0.01, 0.09]	0.03
Negated article	[low] Prior belief < No belief	0.04	[0.00, 0.08]	0.07
Noun	[low] Prior belief < No belief	0.01	[-0.03, 0.06]	0.32
Main verb	[low] Prior belief < No belief	-0.01	[-0.08, 0.05]	0.63

Table 3.5 Regression coefficients.

3.5   Discussion

In the present study, I set out to investigate the incremental interpretation of questions with biases, focusing on questions in German. Crucially, just like in Experiment 1, I addressed this issue using a task in which participants read negated polar questions against discourse contexts which contained different epistemic biases relating to the propositions targeted in the questions. Biases were operationalized as a combination of contextual evidence against the relevant proposition, on the one hand, and the presence or absence of a particular belief pertaining to the proposition, on the other. My predictions were that these biases would interact with the form of a question – more specifically, the syntactic form of the negated

questions – to give rise to differential effects on the measured reading times.

More concretely, I expected reading times at the critical sentence regions – which spanned from the auxiliary verb (*Hast*) to the noun (Hast du nicht ein/ Hast du kein *Ticket*) – to be lower when a high negation question was read in a discourse context containing a particular belief, and similarly so when a low negation question was read in a discourse context containing no particular belief. The hypothesized link between the empirical reading signatures and the underlying processing mechanism states that a particular word is harder to process – and hence takes longer to read – the less predictable that word is in context, in my case predictability being, by design, assumed to be tied to explicit expectations as per the constraints described above.

The quantitative analyses reported in the previous section are mixed with regard to these predictions. In the high negation there was primarily evidence against the prediction, with only weak evidence for the expected effect at the auxiliary, evidence against the expected effect in the critical regions between the auxiliary and the noun, and no reliable difference at the noun. In the low negation, on the other hand, there was no reliable difference at the auxiliary, evidence for the expected effect in the regions between the auxiliary and the noun, and no reliable difference at the noun. Moreover, unlike in Experiment 1, there was no evidence of facilitatory processing in high negation questions at the last sentence region.

All in all, much like in English, I interpret these results as evidence that comprehenders might potentially track biases to do with the epistemic state of a questioner when processing negated polar questions in German. One observation which may reconcile the positive results attested in the low negation with the negative results attested in the high negation has to do with the fact that the tense of the auxiliary construction was often different between *Prior belief* and *No belief* items, as seen in the example items shown on the x-axis of Figure 3.10. While this may have confounded any potential effects of belief type, the decision to design the items using different verb tenses was made to ensure that they sounded as pragmatically appropriate as possible given the constraints of the task. It may very well be that the tense

of the auxiliary served as an additional, if not conflicting, cue with regard to the relevant questioner biases.

3.6 General discussion

3.6.1 Implications for psycholinguistic accounts of pragmatic processing

With regard to what the combined results of the experiments might mean, generally speaking, in terms of pragmatic processing, given the very limited evidence in the data for bias-related pragmatic prediction, it seems reasonable to conclude that, in a task such as the present one, comprehenders do not need to track any biases to do with the epistemic state of a questioner, at least not as they would compared to more naturalistic/ interactive scenarios of language comprehension. Indeed, while other cues derived from the syntactic form of an utterance have been shown to affect online sentence processing in the lab (e.g. Yano and Koizumi, 2018), the biases entertained here do not relate to straightforward referential tracking but rather to more social aspects of language interpretation, such as how a comprehender might *respond* to a question which is biased towards a particular state of affairs, or how an original speaker bias, as discussed in the literature, might affect what a comprehender *perceives* the epistemic/ evidential support to a particular state of affairs to actually be. While response formation was not operationalized in the current study, the findings of the processing tasks seem at odds with the data from the norming tasks, which suggests that the discourse contexts built into the materials do reflect the theoretically-expected biases of interest. However, one should keep in mind that people's offline interpretation preferences, as reflected in the norming data, might not necessarily map onto expectations which are updated incrementally during processing, at least not at the granularity which can be captured using a reaction-time-based method like word-by-word self-paced reading.

Moreover, while negated polar questions do encode a question producer's stance in relation to the expected likelihood of a given response, both in the lab and in the wild these epistemic biases might be secondary in the face of a question's primary pragmatic function – which might be to request information, invite someone somewhere, or request a favor (Heritage and C. Raymond, 2021). And while clear-cut behavioral data with respect to the production of such questions has been shown to be elicitable in the lab (Di Maro et al., 2021; Domaneschi et al., 2017), in online processing the relevant pragmatic biases might only become apparent in light of stronger, potentially more ecologically valid comprehension pressures. Indeed, one seemingly robust finding of recent studies on (probabilistic) predictive processing (e.g. Roettger and Franke, 2019; see also the results from Chapter 2), is that rational processing demands and their accompanying behavioral correlates might not necessarily map onto categorical (Gricean) pragmatic expectations, that is, unless there are strong enough pressures in place such that the relevant expectations, which seem to be probabilistic in nature, actually become resource-rational in context.

All in all, the question of whether epistemic biases can impact incremental question interpretation and the derivation of any potential pragmatic inferences should be further investigated in the lab. A particularly fruitful avenue for future research relates exactly to the conditions under which tracking biases and predicting upcoming material might become pragmatically relevant. In the current study, the experimental task involved reading questions for comprehension, thus not involving any sort of actual contingent response preparation or decision-making. In other words, reading for comprehension might not give rise to enough of a pressure such that comprehenders track potential biases encoded by the form of a question, even if these biases might, in principle, be pragmatically relevant.

Given that polar questions seem particularly well-suited to investigations in terms of (dis)preferred responses, future experimental studies could situate question interpretation not only in more or less felicitous discourse contexts but, crucially too, in pragmatically and/ or otherwise socially accountable action scenarios (see Benz and Gotzner (2020) and

Gisladottir et al. (2015) for examples of such paradigms in experimental pragmatic research). In fact, the empirical observation that the participants in Experiment 1 tracked the beliefs built into the discourse situations when processing high negation questions is evidence that comprehenders are sensitive to pragmatic biases in question interpretation, only that those biases might show in a gradient manner, depending on the relevant pressures at play during comprehension.

3.6.2 Implications for theoretical accounts of questions and pragmatic interpretation

While the current results are at best inconclusive with regard to comprehenders' online sensitives to the pragmatic biases associated with the form of a negated polar question, they do shed some light onto the question of what type of interpretation process might be at play in the processing of questions with biases, and what meaning might be more strongly associated with them.

Regarding the interpretation process potentially underlying the processing of questions with biases, it seems reasonable to assume that, at least in the case of negated polar questions, interpretation might involve some form of inferencing, as argued in, e.g., Romero and Han (2004) and Van Rooy and Safarova (2003) (see also Lauer, 2014). I base that conclusion on the observation that, according to my data, interpreting actual epistemic biases seems not to be directly relevant to the processing of biased questions, at least not as relevant as parsing the utterance and tracking potential referents in discourse. Indeed, recall that, in line with some theoretical accounts and the current empirical picture, the bias in a negated polar question consists of a combination between original speaker bias and contextual evidence. In the present study, I only investigated negative contextual evidence against the proposition. It could be that other types of evidence might make interpreting epistemic biases more relevant when processing biased questions, although the absence of a conclusive result in that

115

regard might have to do, as discussed above, with practical limitations of the experimental paradigm/ materials or the actual granularity of the measurements.

All in all, I'd like to conclude that negated polar questions, as a window into the larger domain of questions with biases and, more generally, pragmatic interpretation, might not have epistemic biases strongly associated with them, at least not such that they impact incremental interpretation in self-paced reading. Instead, it seems that any expectations associated with such questions might require strong contextual demands in order for them to systematically impact online interpretation. Ultimately, epistemic biases might play a non-crucial role in the interpretation of questions, both in terms of strict semantic understanding and more generally in terms of recognizing any stances or pragmatic meanings signaled by questions in discourse. Future research should thus tackle these two facets of biased questions – their epistemic biases and the exact functions they perform in discourse – in an integrated way, exploring the circumstances under which one might affect the other and, perhaps more importantly, the circumstances under which one cannot be disentangled from the other.

REFERENCES

Benz, Anton and Nicole Gotzner (2020). "Embedded implicature: what can be left unsaid?" In: *Linguistics and Philosophy*, pp. 1–32.

Büring, Daniel and Christine Gunlogson (2000). "Aren't positive and negative questions the same?" In: *Unpublished manuscript, UCSC/UCLA. Available online at: http://semantics archive. net/Archive/mYwOGNhO/polar_ questions. pdf.*

Bürkner, Paul-Christian (2017). "brms: An R package for Bayesian multilevel models using Stan". In: *Journal of statistical software* 80.1, pp. 1–28.

Dayal, V. (2016). *Questions*. Oxford Surveys in Semantics an. Oxford University Press. ISBN: 978-0-19-928127-5.

De Ruiter, Jan P (2012). *Questions: Formal, functional and interactional perspectives*. Vol. 12. Cambridge University Press.

Di Maro, Maria, Antonio Origlia, and Francesco Cutugno (2021). "Polar Express: Polar Question Forms Expressing Bias-Evidence Conflicts in Italian". In:

Domaneschi, Filippo, Maribel Romero, and Bettina Braun (2017). "Bias in polar questions: Evidence from English and German production experiments". In: *Glossa: a journal of general linguistics* 2.1.

Farkas, Donka F and Floris Roelofsen (2017). "Division of labor in the interpretation of declaratives and interrogatives". In: *Journal of Semantics* 34.2, pp. 237–289.

Féry, Caroline and Shinichiro Ishihara (2016). *The Oxford handbook of information structure*. Oxford University Press.

Gärtner, Hans-Martin and Beáta Gyuris (2017). "On delimiting the space of bias profiles for polar interrogatives". In: *Linguistische Berichte* 251, pp. 293–316.

Gisladottir, Rosa S., Dorothee J. Chwilla, and Stephen C. Levinson (2015). "Conversation Electrified: ERP Correlates of Speech Act Recognition in Underspecified Utterances". In: *PLOS ONE* 10.3, pp. 1–24. DOI: 10.1371/journal.pone.0120068.

Groenendijk, Jeroen Antonius Gerardus and Martin Johan Bastiaan Stokhof (1984). "Studies on the Semantics of Questions and the Pragmatics of Answers". PhD thesis. Univ. Amsterdam.

Han, Chung-hye (2002). "Interpreting interrogatives as rhetorical questions". In: *Lingua* 112.3, pp. 201–229.

Heritage, John (2012). "Epistemics in action: Action formation and territories of knowledge". In: *Research on language & social interaction* 45.1, pp. 1–29.

Heritage, John and Chase Raymond (2021). "Preference and polarity: Epistemic stance in question design". In: *Research on Language and Social Interaction* 54.1, pp. 39–59.

Jegerski, Jill and Bill vanPatten (2013). *Research methods in second language psycholinguistics*. Routledge.

Kaiser, Elsi and John C Trueswell (2004). "The role of discourse context in the processing of a flexible word-order language". In: *Cognition* 94.2, pp. 113–147.

Karttunen, Lauri (1977). "Syntax and semantics of questions". In: *Linguistics and philosophy* 1.1, pp. 3–44.

Krifka, Manfred (2017). "Negated polarity questions as denegations of assertions". In: *Contrastiveness in information structure, alternatives and scalar implicatures*. Springer, pp. 359–398.

Krifka, Manfred and Renate Musan (2012). "Information structure: Overview and linguistic issues". In: *The expression of information structure*, pp. 1–44.

Kučerová, Ivona and Ad Neeleman (2012). *Contrasts and positions in information structure*. Cambridge University Press.

Kurumada, Chigusa et al. (2014). "Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings". In: *Cognition* 133.2, pp. 335–342.

Ladd, D Robert (1981). "A first look at the semantics and pragmatics of negative questions and tag questions". In: *Papers from the... Regional Meeting. Chicago Ling. Soc. Chicago, Ill.* 17, pp. 164–171.

Lauer, Sven (2014). "Mandatory implicatures in Gricean pragmatics". In: *Formal & Experimental Pragmatics Workshop, ESSLI 2014*, pp. 21–28.

Lee, Seung-Hee (2015). "Two forms of affirmative responses to polar questions". In: *Discourse Processes* 52.1, pp. 21–46.

Mithun, Marianne (2018). "Factors behind variation in marking information structure". In: *Information Structure in Lesser-described Languages: Studies in prosody and syntax* 199, p. 119.

Ozerov, Pavel (2021). "Multifactorial Information Management (MIM): summing up the emerging alternative to Information Structure". In: *Linguistics Vanguard* 7.1.

— (2018). "Tracing the sources of information structure: Towards the study of interactional management of information". In: *Journal of Pragmatics* 138, pp. 77–97.

Raymond, Chase and John Heritage (2021). "Probability and valence: Two preferences in the design of polar questions and their management". In: *Research on Language and Social Interaction* 54.1, pp. 60–79.

Raymond, Geoffrey (2003). "Grammar and social organization: Yes/no interrogatives and the structure of responding". In: *American sociological review*, pp. 939–967.

Roettger, Timo B and Michael Franke (2019). "Evidential strength of intonational cues and rational adaptation to (un-) reliable intonation". In: *Cognitive science* 43.7, e12745.

Romero, Maribel (2020). "Form and Function of Negative, Tag, and Rhetorical Questions". In: *The Oxford Handbook of Negation*.

Romero, Maribel and Chung-hye Han (2004). "On negative yes/no questions". In: *Linguistics and philosophy* 27.5, pp. 609–658.

Searle, John R (1969). *Speech acts: An essay in the philosophy of language.* Vol. 626. Cambridge university press.

Song, Sanghoun (2017). *Modeling information structure in a cross-linguistic perspective.* Language Science Press.

Stenström, Anna-Brita (1984). *Questions and responses in English conversation.* Krieger Pub Co.

Sudo, Yasutada (2013). "Biased polar questions in English and Japanese". In: *Beyond expressives: Explorations in use-conditional meaning.* Brill, pp. 275–295.

Van Rooy, Robert and Marie Safarova (2003). "On polar questions". In: *Semantics and Linguistic Theory.* Vol. 13, pp. 292–309.

Yano, Masataka and Masatoshi Koizumi (2018). "Processing of non-canonical word orders in (in) felicitous contexts: Evidence from event-related brain potentials". In: *Language, Cognition and Neuroscience* 33.10, pp. 1340–1354.

Chapter Four

Interpreting utterances with modal particles and tracking referents in the online processing of referring expressions

4.1 Introduction

In Chapter 3, I presented a study focused on whether comprehenders predictively process utterances containing morphosyntactic cues to epistemic stance. More specifically, I investigated whether comprehenders draw predictive inferences about upcoming lexical information when processing negated polar questions in German and English with different morphosyntactic forms. Crucial for my analysis was the assumption that different forms of a negated polar question – such as, in English, *Didn't you buy a ticket?* vs. *Did you not buy a ticket?* – signal different epistemic meanings, either a commitment to the truth of the proposition expressed in the question in the case of the former or no commitment to that same state of affairs in the case of the latter. In the present chapter, I present a study focused on a different class of modal markers, namely lexical rather than morphosyntactic markers of stance.

In contrast to the morphosyntactic cues investigated in Chapter 3, here I investigate

German adverbs which serve a specialized function as modal particles, i.e., as markers of common ground relations and intersubjective stance. I start by reviewing previous work on the function and interpretation of discourse particles, with a particular focus on particles which serve modal functions like German *eigentlich* and *tatsächlich*, before turning to the empirical study which makes up the present chapter.

4.1.1 Discourse particles and the marking of modality

Discourse markers are pervasive communicative devices found across the languages of the world (Heine et al., 2021). As their name indicate, these are devices used to manage linguistic communication and interaction itself (Fedriani and Sansó, 2017; Jucker and Ziv, 1998; Maschler and Schiffrin, 2015), covering a multitude of functions and meanings, from more subjective, attitudinal ones to fundamentally intersubjective, interaction-oriented ones. Not only do discourse markers, as a class, cover a wide range of functions, individual discourse markers themselves also tend to be polyfunctional, their exact meaning depending on both the utterance they occur in as well as the larger discourse context (see refereces above), as in the examples below which show the flexibility of the English marker *really*.

A: Did you see how fast that bird flew by?

(1) B: I *really* don't think it flew by that fast.

(2) B: I don't *really* think it flew by that fast.

(3) B: I don't think it flew by that fast, *really*.

A: Did you see how fast that bird flew by?

(4) B: *Really*? I didn't even notice.

(5) B: I don't *really* think it flew by that fast.

(6) B: It *really* did fly by fast, you're right.

Really, while being formally an adverb, serves a variety of different discourse-marking functions in the examples above: in (1)-(3) it acts as a marker of (subjective) epistemic stance, indicating, in each case, B's degree of commitment to the truth of the state of affairs at issue in the sequence; in (4) it acts as a surprise/ mirative marker, prefacing B's actual response to A's statement; in (5), much like in (1)-(3), it acts as a marker of subjective stance, indicating B's disagreement to A's statement; finally, in (6), it acts as an intensifier, indicating, respectively, both B's subjective and intersubjective understanding of A's statement.

Despite different meanings being encapsulated in a single form in the case of *really*, discourse markers can have a much more restricted meaning profile depending, among other things, on the lexical category they stem from. Indeed, while *really* formally consists of an adverb, discourse markers can stem from all sorts of lexical categories, from individual words (e.g. *really*) and phrases (e.g. *in fact*) to entire sentence frames (e.g. *What surprises me*) and even non-lexical expressions (e.g. confirmation/ hesitation markers like *mhm* or *huh*). Notwithstanding the diversity of forms or even the exact linguistic status of certain discourse markers, many accounts distinguish discourse *markers* more generally from the more restricted class of discourse *particles* (see Abraham (2016) and Fischer (2006), *inter alia*, for discussions on the terminological issue of calling an item a discourse particle vs. a discourse marker). Unlike other linguistic markers, particles are usually understood as non-inflected elements which can occur in isolation and which impart meaning to other linguistic elements, from individual lexical items to entire sentences or utterances. Discourse particles can therefore be distinguished from discourse markers more generally on the basis of their form, such that uninflected lexical elements like *really* and *in fact* might be classified as particles while larger morphosyntactic constructions like *What surprises me* might not[1].

---

[1]Although non-lexical expressions like *mhm* or *huh* might traditionally not be considered particles or even linguistic devices in their own right, from a functional perspective, they form a continuum with other lexicalized and grammaticalized markers.

While this grants discourse particles, by definition, a high degree of meaning flexibility relative to non-lexical discourse markers, their function as discourse-managing devices is, opposed to their morphological status, their most defining feature.

In the present study, I focus my attention on discourse particles which are used to mark modal relations related to epistemicity. Given their function as markers of modality, such particles are also known in the literature as modal particles, although their exact relation to discourse particles, and by extension discourse markers, is the subject of much debate (see, e.g., Abraham (2016) and Schoonjans (2013)). Indeed, both discourse particles in general and modal particles in particular function as indexical linguistic devices, not modifying the content of an utterance but rather adding to its illocutionary dimension and relating it to other information in the discourse context. However, unlike discourse particles which primarily relate their host utterance to non-propositional information, such as in the English example in (7) below, modal particles serve a more specialized function of relating an utterance to a proposition or an speech-act alternative, as in the German example in (8) below.

(7)   A: Did you see how fast that bird flew by?

      B: *Right*, but that's not really surprising. It was a swallow, and, as you know, they always fly so fast.

(8)   A: Hast du gesehen, wie schnell der   Vogel vorbeigeflogen ist?
         Have you seen   how fast   SING.DEF bird   flew by   is?
         "Did you see how fast that bird flew by?"

      B: Ja, aber verwunderlich ist das nicht. Das   war *ja* eine Schwalbe, die
         Yes, but surprising   is that not.   That was *ja* a   swallow   PL.DEF
         fliegen immer so schnell.
         fly   always so fast.
         "Right, but that's not really surprising. It was, *after all*, a swallow, and they always fly so fast."

In (7), the particle *right* is used to mark a general sense of agreement to the previous

turn while also signaling a transition of turns between A and B. While it might be said to confirm the state of affairs at issue in the sequence, i.e., the speed with which the bird flew by, its primary function is that of a response particle, much like its German counterpart *ja* at the beginning of B's response in (8). Both *right* and the turn-initial *ja* serve as overt confirmation of the at issue state of affairs expressed in the statement by A. The *ja* in the second utterance in (8), on the other hand, plays a similar yet still qualitatively different role: it signals that the utterance it is embedded in is uncontroversial, doing so by establishing that the proposition expressed in the utterance is part of the common ground between A and B, i.e., that they both know that the bird might have been a swallow, as suggested by the approximate English translation of the utterance-medial *ja*. While both *right* in the English example and the first *ja* in the German example thus function primarily as discourse-managing devices – more specifically response or expectation-managing devices – facilitating the flow of the conversation while indirectly also contributing to the update of shared epistemic information, the second *ja* in the German example, a so-called modal particle, is used explicitly to manage common ground between the interlocutors, not serving as a response particle in itself and only being felicitous because it's part of an utterance which has propositional content of its own.

Despite discourse particles being, cross-linguistically, a common means of marking relations of knowledge and truth and any uncertainty surrounding them (Heine et al., 2021), the two examples above illustrate the fact that languages differ in terms of the exact types of particles they employ in the expression of such relations. The empirical picture is such that while some languages, like English, draw mostly on manner and certainty adverbs (e.g. *really*, *right*, *apparently*), other languages, like German, rely more heavily on particles with a specialized modal function. Focusing on the modal-marking repertoire of German, there are a variety of particles dedicated primarily to the expression of epistemicity, for instance, *ja*, *doch*, and *eigentlich*, to name just a few. While all of these particles can be used to establish a relation of knowledge or belief between two propositions, they do so in different

ways. Let's look at the examples below.

A: Hast du gesehen, wie schnell der   Vogel vorbeigeflogen ist?
    Have you seen   how fast   SING.DEF bird   flew by   is?
    "Did you see how fast that bird flew by?"

(9)   B: Ja,   aber verwunderlich ist das   nicht. Das   war *ja* eine Schwalbe, die
        Yes, but   surprising   is   that not.   That was *ja* a   swallow   PL.DEF
        fliegen immer so schnell.
        fly   always so fast.
        "Right, but that's not really surprising. It was, *after all*, a swallow, and they
        always fly so fast."

(10)   B: Der   ist *doch* nicht so schnell vorbeigeflogen.
        SING.DEF is   *doch* not   so fast   flew by.
        "It didn't *actually* fly by that fast."

In (9), the utterance-medial *ja* can be said to mark B's commitment to the truth of the proposition *Das war eine Schwalbe*, or to put it differently, it can be said to indicate their belief in the strength or validity of that particular state of affairs (Döring and Repp, 2019; Gast, 2008; Unger, 2016), which is pragmatically reinforced by the ensuing statement "and they always fly by so fast". More specifically, as discussed with regard to the example in (8), *ja* indexes that B believes that the proposition at hand is part of the common ground between them and their interlocutor, such that that particular piece of information is almost trivial or unremarkable when uttered (Döring and Repp, 2019; Gast, 2008; Unger, 2016). As such, *ja* can be said to mark not only B's *subjective* stance toward the relevant state of affairs (i.e. "I commit to the truth of that being a swallow") but also an *intersubjective* understanding of how their response relates to A's original statement (i.e. "I think we both agree that that's not surprising given that the bird was a swallow and they always fly so fast"). *Doch* functions in a very similar way to *ja*[2], except that it indexes a mismatch of

---

[2]Note, however, that *doch* is more flexible than *ja*, occurring both sentence-medially as well as sentence-initially as an intensifier/ focus particle (*Doch ist der schnell vorbeigeflogen, die fliegen immer so schnnell*; English: "*Of course* it flew by fast, they always fly really fast").

information in the common ground, indicating that something is in fact false while it was originally believed to be true (Döring and Repp, 2019; Gast, 2008; Unger, 2016). Thus, while A states that they think that the bird flew by fast, in (10) B challenges that position stating that the bird didn't actually fly by that fast. Let's now compare the usage of *ja* and *doch* to that of two other similar particles, *eigentlich* and *tatsächlich*.

A: Hast du gesehen, wie schnell der Vogel vorbeigeflogen ist?
Have you seen how fast SING.DEF bird flew by is?
"Did you see how fast that bird flew by?"

(11) B: Der ist *tatsächlich* sehr schnell vorbeigeflogen.
SING.DEF is *tatsächlich* very fast flew by.
"It did *actually* fly by very fast." OR "It *really* did fly by very fast."

(12) B: Der ist *eigentlich* nicht so schnell vorbeigeflogen.
SING.DEF is *eigentlich* not so fast flew by.
"It didn't *actually* fly by that fast."

Tatsächlich in (11) serves a function similar to that of *ja* in (8) and (9): it confirms some piece of information which is part of the common ground between the two interlocutors. Unlike *ja*, however, it confirms that the proposition expressed in the statement by A is true without indexing that that information is trivial or unremarkable. It therefore plays a similar role to that of a positive response particle like *yes* or *right*; just like *ja*, however, it is only felicitous because it's part of an utterance which has propositional content of its own. In keeping with the parallelisms between particles, *eigentlich* is to *doch* very much like *tatsächlich* is to *ja*, that is, it disconfirms some information which is part of the common ground between interlocutors. More specifically, in (12), *eigentlich* disconfirms the statement that the bird flew by really fast, contrasting that proposition with the proposition that the bird didn't actually fly by that fast. *Eigentlich* therefore introduces a new proposition which stands in direct contrast to the proposition originally expressed in the statement by A, playing a similar role to that of a negative response particle like *no* while, just like *tatsächlich*, being dependent on the utterance it modifies.

In summary, *ja*, *doch*, *tatsächlich*, and *eigentlich* are all instances of discourse particles in German which serve the specialized function of relating the utterance they are embedded in to a proposition or speech-act alternative which may be more or less salient in the discourse context. Despite their common function as epistemicity markers, these particles differ from one another in terms of the specific modal meaning they index: while *ja* and *tatsächlich* both mark agreement between the propositional content of their host utterance and some contextually-relevant propositional antecedent, *doch* and *eigentlich*, on the other hand, both mark contrast. Interestingly, what the examples above show is that these particles are polyfunctional by definition: even in the same context of utterance, they can be said to have both a more propositionally-oriented function, which makes them modal particles in the strict sense, as well as a discourse-managing function, which brings them together with other discourse-managing devices like response particles. Crucially, when it comes to how these particles are interpreted in context, interpretation seems to depend on a pragmatic calibration between their specialized modal meaning and their general discourse-marking meaning. As discussed in Chapter 1, conventional implicatures, when construed as a form of pragmatic inference (Feng, 2011, 2010), can be thought to capture exactly such process, that is, the derivation of a context-sensitive interpretation by means of association between contextual constraints and the conventionalized meaning(s) of an expression.

In the study reported below, I address the question of whether modal particles, as discussed so far, serve as predictive cues to interpretation in the incremental processing of language, as measured in a mouse tracking task. More specifically, I focus on the interpretation of referring expressions containing the German particles *tatsächlich* and *eigentlich*. Given the similar and yet complementary functions of *tatsächlich* and *eigentlich*, I'm interested in whether processing one or the other particle gives rise to differential effects on real-time interpretation and the resolution of reference. Before introducing the study and reporting my findings, I review previous findings on the processing of modal particles and related non-lexical cues of epistemicity.

4.1.2 Processing profile of modal particles

Previous psycholinguistic research has explored the role of different epistemicity markers in the incremental processing of language. Indeed, many studies concerned with the role of information structure on sentence processing have looked at the processing of cues which can be said to function as non-lexical markers of epistemicity. For example, Kurumada et al. (2014) looked at whether contrastive prosody in English allow comprehenders to predict the likely referent in referring expressions of the sort "It looks like a zebra" when either the verb (i.e. "It LOOKS like a zebra") or the noun (i.e. "It looks like a ZEBRA") is accented. In order to test that, they had participants listen to the expressions while looking at visual displays containing images of potential referents, their eye movements being tracked in the process. Participants were instructed to select the most likely referent by clicking on the relevant image. Crucially, the visual displays contained not only the mentioned referent (e.g. a zebra) but also a perceptually similar competitor (e.g. an okapi) and two distractors. Kurumada et al. (2014) found that a contrastive accent on the verb, as opposed to the noun, induced comprehenders to look predictively at the non-prototypical referent, i.e., something that looked like a zebra but wasn't one, prior to the onset of the noun.

In a similar study, Roettger and Franke (2019) looked at whether contrastive prosody in German allow comprehenders to draw predictive inferences about the likely referent in referring expressions of the sort "Der Wuggy hat dann die Birne aufgesammelt" (English: the wuggy has then the pear picked-up, "The wuggy then picked up the pear"). Crucially, they manipulated whether the auxiliary verb or the object of the sentence was accented. Much like in the study by Kurumada et al. (2014), they had participants listen to the expressions while looking at a visual displays containing images of potential referents. Participants were instructed to select the most likely referent by clicking on the relevant image, their mouse movements being tracked in the process. Roettger and Franke (2019) found that a pitch accent on the auxiliary (i.e. "Der Wuggy HAT DANN die Birne aufgesammelt"; English:

the wuggy HAS THEN the pear picked-up) induced comprehenders to shift their mouse towards the mentioned referent earlier compared to when the pitch accent fell on the object of the sentence (i.e. "Der Wuggy hat dann die BIRNE aufgesammelt"; English: the wuggy has then the PEAR picked-up).

Taken together, these studies suggest that an intonational feature like pitch accent can guide comprehenders in the online resolution of reference, serving as a cue to intended meaning in context. While prosody can thus function as a non-lexical means of marking epistemic stance in reference resolution, at least when used in combination with other linguistic markers like the expression "It looks like" as in Kurumada et al. (2014), little research to date has directly addressed the processing of lexical markers of epistemicity like discourse and modal particles. Indeed, only four studies in the literature have explicitly looked at the processing of modal particles. Dörre, Czypionka, et al. (2018) investigated whether particles in German which function both as modal and non-modal markers are processed differently depending on the relative frequency of each meaning and on whether the discourse context biases either one or the other interpretation. They tested 10 different particles, having participants read sentences of the sort "Wer hat *bloß* den Flur gewischt?" (English: who has *bloß* the corridor wiped; "Who *bloß* wiped the corridor?") one word at a time against contexts which disambiguated between the particles' modal and non-modal meanings, as in 13 and 14 below.

(13) **Modal usage**

Hier ist noch Schneematsch von draußen, wer hat *bloß* den Flur   gewischt?
Here is still mud   from outside   who has *bloß* the corridor wiped?

"Here there's still some mud from outside, who *on earth* wiped the corridor?"

(14) **Non-modal usage**

Die anderen Zimmer sind ebenfalls schmutzig, wer hat *bloß* den Flur
The other   rooms   are also   dirty   who has *bloß* the corridor
gewischt?
wiped?

"The other rooms are also dirty, who wiped *only* the corridor?"

Crucially, Dörre, Czypionka, et al. (2018) measured the reading times at each word in the sentence, their analysis including a measure of the relative meaning frequency of each particle, as computed from a corpus of spoken German. They found that comprehenders read the particle itself at the same speed regardless of the type of context and the relative frequency of its different meanings. One word downstream from the particle (i.e. "Wer hat bloß *den* ..."), while the reading times did not differ depending on the context, they were reliably higher when the processed particle had a more frequent modal meaning. Two words downstream from the particle (i.e. "Wer hat bloß den *Flur* ..."), the reading times were reliably higher when the context biased a modal interpretation as opposed to a non-modal one. Finally, at the last word in the sentence (i.e. "Wer hat bloß den Flur *gewischt*"), three words downstream from the particle, while the reading times were generally shorter when the processed particle had a more frequent modal meaning as well as when the context biased that same meaning, they were longer when considering more frequent modal meanings in modal-biasing contexts. All in all, the results suggest that both context and the relative frequency of a particle's modal meaning impact the online processing of sentences containing modal particles, such that downstream from the particle processing is, by and large, more costly the more conventionalized its modal meaning. Interestingly, while a modal-biasing context and a high relative modal meaning frequency seem to facilitate processing at the main verb in the sentence when considered separately, their interaction actually seems to hinder processing, suggesting that, overall, the modal meaning of a polyfunctional particle is harder to process than its non-modal meaning.

In a subsequent study, Dörre and Trotzke (2019) further investigated the processing of two of the particles originally investigated in Dörre, Czypionka, et al. (2018), namely *bloß* and *nur*. In their modal usage, both *bloß* and *nur* index a (subjective) strengthening of a proposition[3], as in "Warum hat sie *bloß* das Kleid gekauft?" (English: why has she *bloß* the

---

[3]See Dörre and Trotzke (2019) for their account on how the modal meaning of *bloß* and *nur* might relate

dress bought), which can be interpreted as "Why *on earth* did she buy the dress?". *Bloß* and *nur* can also be used as focus particles, functioning like the English particle *only*, as in "Warum hat sie *bloß* das Kleid gekauft?" (English: why has she *bloß* the dress bought), meaning "Why did she *only* buy the dress?". As in Dörre, Czypionka, et al. (2018), Dörre and Trotzke (2019) had participants read questions of the sort "Warum hat sie *bloß* das Kleid gekauft?" one word at a time against contexts which disambiguated between the particles' modal and non-modal meanings. Despite measuring reading times at each individual word in the sentence, they analyzed the entire region starting at the particle and ending at the last word in the sentence (i.e. *particle* + das + Kleid + gekauft). They found that both in the case of *bloß* and *nur* comprehenders read the aggregated critical region at the same speed regardless of whether the context biased the particle's modal meaning or its meaning as a focus marker.

In a second experiment, Dörre and Trotzke (2019) looked at whether prosodic focus affects the processing of *bloß* and *nur*, having participants listen to the same questions as in their reading experiment and manipulating whether the noun (i.e. "Warum hat sie *bloß* das KLEID gekauft?") or the auxiliary verb (i.e. "Warum HAT sie *bloß* das Kleid gekauft?") was accented, just like in the studies by Kurumada et al. (2014) and Roettger and Franke (2019). Crucially, while the focus-marking function of *bloß* and *nur* is said to be realized prosodically via a pitch accent on the noun (i.e. the focused element), their modal-marking function does not strongly correlate with any particular intonational contour, either the question form, the auxiliary verb, or the noun being accented in such a usage. Participants were instructed to listen to the test sentences, presented in the absence of any discourse context. After listening to the sentences, they were presented with the same discourse contexts used the reading experiment, namely one biasing a modal interpretation (e.g. "Hier ist noch Schneematsch von draußen"; English: Here is still mud from outside, "Here there's still some mud from outside") and one biasing a non-modal interpretation

to the particles' meaning as focus particles.

131

(e.g. "Die anderen Zimmer sind ebenfalls schmutzig"; English: The other rooms are also dirty, "The other rooms are also dirty"). Participants were then asked to select the context that was most appropriate. Dörre and Trotzke (2019) found that when comprehenders were presented with sentences where the auxiliary was accented they selected the modal-biasing context at a reliably higher rate (87%) than the non-modal-biasing context (13%), whereas when they were presented with sentences where the noun was accented they selected the non-modal-biasing context at a reliably higher rate (63%) than the modal-biasing context (37%). The results suggest that contrastive prosody acts as a reliable cue in the processing of modal particles, guiding comprehenders in their interpretation of polyfunctional particles such as *bloß* and *nur*.

The two other studies in the literature which have looked at the processing of modal particles have dealt with the particles *inderdaad* and *eigenlijk* in Dutch. These two particles have the same meaning profiles as *tatsächlich* and *eigentlich* in German, respectively, meaning that *inderdaad* marks agreement between two discourse elements while *eigenlijk* marks contrast. Much like their counterparts in German, they also serve both a more specialized modal function and a general discourse-managing function. While in their discourse-managing function *inderdaad* and *eigenlijk* mark a general sense of (dis)agreement between two discourse elements, like a response particle would, in their modal function they mark (dis)agreement between the propositional content of their host utterance and some contextually relevant propositional antecedent.

van Bergen and Bosker (2018) investigated whether *inderdaad* and *eigenlijk* allow comprehenders to draw predictive inferences about the likely referent in referring expressions of the sort "Ik schrok *inderdaad* van de rondrennende *BEEP* aan het eind" (English: I startled *inderdaad* from the running *BEEP* at the end; "I was indeed scared by the running *BEEP* at the end"). The referring expressions were embedded in mini dialogues, as in (15) below.

(15) Ondanks haar angst voor dieren   is Marie naar het circus geweest.
      Despite her   fear   from animals is Marie to   the circus gone.

      "Despite her fear of animals, Marie went to the circus."

      Je   vond de dierenact zeker doodeng?
      You found the animal act surely terrifying?

      "You must have been terrified by the animal act?"

      Ik schrok *inderdaad* van de rondrennende *BEEP* aan het eind.
      I   startled *inderdaad* from the running   *BEEP* at   the end.

      "I was *indeed* scared by the running *BEEP* at the end."

Participants were instructed to listen to the dialogue while looking at visual displays containing images of potential referents, their task being to select the most likely referent by clicking on the relevant image. Their eye movements were tracked in the process. Crucially, the displays contained a referent which was biased by the discourse context as being the target of the referring expression (e.g. a lion in (15)) as well as a contextually-relevant competitor (e.g. a clown) and two distractors. While a modal interpretation of *inderdaad* should lead to a confirmation of the proposition expressed in "You must have been terrified by the animal act?", resulting in a preference for expecting the referent biased by the context (i.e. the lion), a modal interpretation of *eigenlijk* should lead to a disconfirmation of the proposition expressed in "You must have been terrified by the animal act?", resulting in a preference for expecting the unbiased referent (i.e. the clown.) By design, however, the actual referent was beeped in the playback of the target sentence, as indicated in (15). The omission of the critical noun was meant to avoid biasing a strictly modal interpretation of the particles, such that participants could interpret them not necessarily in propositional terms but perhaps in discourse-managing terms more generally. While it is unclear what referent in the visual display could be mapped onto a non-modal interpretation of *inderdaad,* a non-modal interpretation of *eigenlijk* could induce comprehenders to interpret it as marking a general sense of disagreement between its host utterance and the preceding context, thus not strongly supporting a referential contrast between "You must have been terrified by the

133

animal act" and "I was scared by the running *BEEP* at the end", which should lead to no particular preference for the unbiased referent.

van Bergen and Bosker (2018) found that processing *inderdaad* induced comprehenders to select the referent biased by the context (i.e. the lion) at a reliably higher rate (about 100%) than the unbiased competitor (i.e. the clown). They also found that it induced comprehenders to direct their gaze at the biased referent prior to the onset of the beep at a reliably higher rate, such that, at the beep onset, about 50% of the eye fixations were directed at the biased referent while about 20% of them were directed at the unbiased referent. When it comes to *eigenlijk*, van Bergen and Bosker (2018) found that processing the particle induced comprehenders to select the unbiased referent (i.e. the clown) at a reliably higher rate (about 60%) than the biased referent. They also found that, when this was the case, it induced comprehenders to direct their gaze at the unbiased referent prior to the onset of the beep at a reliably higher rate, such that, at the beep onset, about 40% of the eye fixations were directed at the unbiased referent while about 20% of them were directed at the biased referent. When comprehenders selected the biased referent after processing *eigenlijk*, they did not direct their gaze at the unbiased referent prior to the onset of the beep at a reliably higher rate, such that, at the beep onset, about 25% of the eye fixations were directed at the unbiased referent while about 35% of them were directed at the biased referent.

In a second experiment, van Bergen and Bosker (2018) extended their original design by introducing a new *eigenlijk* condition where the discourse context did not bias any particular referent, as in (16) below.

(16) Op zaterdagmiddag   is Marie naar het circus geweest.
      On Saturday afternoon is Marie to   the circus gone.
      "On Saturday afternoon, Marie went to the circus."

      Wat   is je   het best bijgebleven?"
      What is you the best remembered?
      "What did you find most memorable?"

> Ik schrok *eigenlijk* van de rondrennende *BEEP* aan het eind.
> I   startled *eigenlijk* from the running   *BEEP* at   the end.
>
> "I was *actually* scared by the running *BEEP* at the end."

They predicted that this would result in no particular referential expectation compared to the *eigenlijk* condition where the context biased one of the referents. As for the remaining conditions, they did not predict any changes compared to the first experiment. Indeed, just like in their first experiment, van Bergen and Bosker (2018) found that processing *inderdaad* induced comprehenders to select the biased referent (i.e. the lion) at a reliably higher rate (about 100%) than the unbiased referent. They also found that it induced comprehenders to direct their gaze at the biased referent prior to the onset of the beep at a reliably higher rate, such that, at the beep onset, about 55% of the eye fixations were directed at the biased referent while about 15% of them were directed at the unbiased referent. When it comes to *eigenlijk* in the biasing condition, van Bergen and Bosker (2018) found that processing the particle did not induce comprehenders to select the unbiased referent (i.e. the clown) at a reliably higher rate than the biased referent. However, when comprehenders selected the unbiased referent, they did direct their gaze at the unbiased referent prior to the onset of the beep at a reliably higher rate, such that, at the beep onset, about 45% of the eye fixations were directed at the unbiased referent while about 25% of them were directed at the biased referent. On the other hand, when comprehenders selected the biased referent after processing *eigenlijk* in the biasing condition, they did not direct their gaze at the unbiased referent prior to the onset of the beep at a reliably higher rate, such that, at the beep onset, about 25% of the eye fixations were directed at the unbiased referent while about 40% of them were directed at the biased referent. Interestingly, when it comes to *eigenlijk* in the unbiasing condition, van Bergen and Bosker (2018) again found that processing the particle did not induce comprehenders to select the unbiased referent (i.e. the clown) at a reliably higher, much like in the biasing *eigenlijk* condition. They also found that, when comprehenders selected the biased referent, it did not induce them to direct their gaze at

the biased referent prior to the onset of the beep at a reliably higher rate, such that, at the beep onset, about 40% of the eye fixations were directed at the biased referent while about 25% of them were directed at the unbiased referent.

All in all, the results suggest that processing modal particles affects the interpretation of referring expressions and the online resolution of reference. While processing an agreement-marking particle like *inderdaad* induces comprehenders to expect an agreement between its host utterance and a contextually-relevant antecedent, allowing them to predict a referent biased by the context, processing a contrast-marking particle like *eigenlijk* induces comprehenders to expect a disagreement between its host utterance and a contextually-relevant antecedent, allowing them to predict a contextually-relevant referential alternative. Interestingly, *eigenlijk* does not seem to be biased in terms of either a modal or a non-modal interpretation, regardless of whether the discourse context biases a modal interpretation. However, when *eigenlijk* is interpreted as marking a modal relation, it affects online interpretation to a similar extent as *inderdaad.*

In a subsequent study, Rasenberg et al. (2019) investigated whether *inderdaad* and *eigenlijk* allow comprehenders to draw predictive inferences about upcoming lexical information in sentences of the sort "We zijn *inderdaad* elke dag naar een museum geweest" (English: We are *inderdaad* every day to a museum been; "We *indeed* went to a museum every day"), which were embedded in mini dialogues, as in (17) below.

(17) Diana is een weekend met haar klas van   de kunstacademie naar Parijs
      Diana is a   weekend with her   class from the art school   to   Paris
      geweest.
      been.
      "Diana spent a weekend in Paris with her art academy class."

      Haar vriendin vraagt: Jullie hebben vast een hoop kunst gezien?
      Her   friend   asks: You   have   plenty of   art   seen?
      "Her friend asks: You guys must have seen a lot of art?"

      Diane zegt: We zijn *inderdaad* elke   dag naar een **museum**/ **park** geweest.
      Diana says: We are *inderdaad* every day to   a   **museum**/ **park** been.

136

"Diana says: We *indeed* went to a **museum**/ **park** every day."

Participants were instructed to read the dialogue for comprehension, their event-related potentials (ERPs) being measured relative to the words in the sentences. Crucially, Rasenberg et al. (2019) compared the ERP signatures relative to the noun following the modal particle (bolded in (17)), which was either predictable (e.g. museum) or unpredictable (e.g. park) given the discourse context. In order to draw inferences about differences in the amplitude of the relevant ERP components, they compared the signal for each particle to that of a control condition containing adverbs instead of discourse particles, as in "We zijn *daar* elke dag naar een museum geweest" (English: We are *there* every day to a museum been; "We went *there* to a museum every day"). They predicted qualitative changes to the amplitude of the N400 and P600 components, the former functioning as an index of semantic plausibility and the latter as an index of prediction disconfirmation. Whereas upon processing *inderdaad* the amplitude of the N400 component should be reduced for predictable nouns (e.g. museum) and increased for unpredictable nouns (e.g. park) relative to processing an adverb, upon processing *eigenlijk* the amplitude of the N400 component should be reduced for unpredictable nouns and increased for predictable nouns. Similarly, whereas upon processing *inderdaad* the amplitude of the P600 component should be decreased for predictable nouns and increased for unpredictable nouns relative to processing an adverb, upon processing *eigenlijk* the amplitude of the P600 component should be decreased for predictable nouns and increased for unpredictable nouns. Rasenberg et al. (2019) found that processing neither *inderdaad* nor *eigenlijk* affected the N400 component of the ERP signal. Similarly, they found that neither particle affected the P600 component.

Taken together, the studies in the literature paint a complex picture of the processing profile of modal particles. While the earlier results from van Bergen and Bosker (2018) suggest that processing modal particles does affect incremental interpretation, such that comprehenders are able to draw predictive inferences about subsequent material in a sentence when

that material is relevant for interpretation, the more recent results from Rasenberg et al. (2019) do not provide evidence that modal particles allow comprehenders to draw predictive inferences about linguistic material downstream in a sentence. The results from Dörre, Czypionka, et al. (2018) further suggest that processing modal particles is at least partially more costly the more conventionalized the modal meaning associated with a particular particle form. These findings raise a number of questions about the role of modal particles as cues to epistemicity in the incremental processing of language.

First and foremost, given the conflicting results, it remains unclear to what extent modal particles affect incremental interpretation. While they do affect the outcome of the interpretive process, as shown in van Bergen and Bosker (2018), interacting with the propositional meaning of the sentences they occur in in expected ways given their modal meaning, it is not clear what precise role the non-sentential discourse context plays in biasing any relevant interpretitve expectations. For instance, in the study by van Bergen and Bosker (2018), the same discourse context constrained interpretation qualitatively differently depending on the particle at hand, as *eigenlijk* was interpreted both as a modal marker and as a non-modal discourse-managing particle while *inderdaad* was not. While this was a feature of the experimental design, allowing the authors to tackle the polyfunctionality of *eigenlijk* and *inderdaad*, it obscures the inferences one may draw about the interaction between context and particle meaning. Even when a particular meaning of a particle might be said to be constrained in context, as in Dörre and Trotzke (2019), potentially overriding any concerns about non-trivial interactions between context bias and particle meaning, it remains unclear to what extent modal particles serve as reliable cues to interpretation in context, especially when considering their role as *predictive* cues to interpretation. In the present study[4], I tackle precisely that question, focusing on the incremental interpretation of sentences containing the German modal particles *eigentlich* and *tatsächlich*. Crucially, I investigate the

---

[4]This study has been conceived and carried out in collaboration with Michael Franke, Laine Stranahan, and Timo Roettger.

processing of referring expressions of the sort "Das ist *eigentlich/ tatsächlich* ein Bild von einem Wolf" (English: This is *eigentlich/ tatsächlich* a picture from a wolf; "That's *actually/ indeed* a picture of a wolf") against discourse contexts which bias not only concrete referential expectations but also clear mappings between those referential expectations and a modal interpretation of the respective particles. Thus, I manipulate the predictability of any given referent as a function of an interaction between the discourse context and the usage of the particles in context. I also manipulate whether the particles are used reliably according to their modal meaning. In other words, I look at whether pragmatically incoherent usages of *eigentlich* and *tatsächlich* lead to qualitative modulations of the processing signatures compared to the pragmatically coherent usages. In the section below, I describe the exact predictions of the study, as they relate to the test items, as well as the details and setup of the experimental task.

4.2 Experiment

Method

Participants

Fifty native speakers of German were recruited among the cognitive science student population of the Osnabrück University. Participants were given course credit for their participation in the experiment. Data collection was conducted entirely online and all collected data was stored at servers from the Osnabrück University.

Materials and design

Participants were asked to listen to 120 miniature dialogues mimicking an interaction between a parent and a child, each anchored to a visual world display containing two pictures, one depicting a referent explicitly mentioned in the dialogue and another depicting a semantically

and/ or perceptually-related competitor, e.g., a wolf and a dog, as shown in Figure 4.1. Each dialogue consisted of a question-answer pair where the question, posed by the child, was polar in nature and made direct reference to one of the pictures visible on screen (e.g. "Ist das ein Bild von einem Wolf?"; English: "Is this a picture of a wolf?"), while the answer, provided by the parent, either confirmed or disconfirmed the truth of the proposition expressed in the question. Crucially, the answer did not contain any response or negation particle but rather a discourse particle with a modal meaning, either *eigentlich* or *tatsächlich* (e.g. "Das ist *eigentlich*/ *tatsächlich* ein Bild von einem Wolf"; English: "That's indeed/ actually a picture of a wolf"). Participants' task was to listen to the question-answer sequence and to select the referent they thought was most representative of the topic of the conversation.

Each participant was assigned to one of two experimental groups with a different distribution of critical trials. In the **reliable** group, participants encountered 60 fillers and 60 dialogues in which the discourse particles were used according to their conventionalized modal meaning, i.e., where *eigentlich* marked a contrast and *tatsächlich* marked a confirmation, as shown below.

Question

    Ist das ein Bild   von   einem **Wolf**?
    Is this a   picture from a   **wolf**?

    "Is this a picture of a **wolf**?"

Answer

    Das ist *eigentlich* ein Bild   von   Das ist *tatsächlich* ein Bild   von
    This is *eigentlich* a   picture from   This is *tatsächlich* a   picture from
    einem **Hund**.   einem **Wolf**.
    a   **dog**.   a   **wolf**.

    "That's actually a picture of a **dog**."   "That's indeed a picture of a **wolf**."

In the **unreliable** group, participants encountered 20 fillers, 60 dialogues in which the particles were used according to their conventionalized modal meaning, and 40 dialogues in

which the particles were used unreliably, i.e., where *eigentlich* marked a confirmation and *tatsächlich* marked a contrast, as shown below.

Question

Ist	das	ein	Bild		von	einem	**Wolf**?
Is	this	a	picture	from	a		**wolf**?

"Is this a picture of a **wolf**?"

Answer

Das	ist	*eigentlich*	ein	Bild	von
This is	*eigentlich*	a	picture from		

einem **Wolf**.
a **wolf**.

"That's actually a picture of a **wolf**."

Das	ist	*tatsächlich*	ein	Bild	von
This is	*tatsächlich*	a	picture from		

einem **Hund**.
a **dog**.

"That's indeed a picture of a **dog**."

Filler trials contained various sorts of questions, such as polar questions similar to the ones from critical trials (e.g. "Ist das ein Fisch?"), polar questions with miscellaneous forms (e.g. "Findest du das Tier mit der dicken Nase auch lustig?", "Ist das, etwas was man zum Kämpfen benutzt?"), as well as wh-questions (e.g. "Wie heißt das orangene Tier?"; "Was ist das mit dem Schabel für ein Tier?"). The respective answers to these questions did not contain any discourse particles.

Procedure

Written instructions were provided prior to the actual task, followed by 10 practice trials which mimicked the exact procedure of the test trials. Participants were instructed to listen to the question-answer sequence and to select the picture that best matched the topic of the exchange. On each trial, they were presented with a "Start" button at the center bottom of the screen. Upon clicking on it, the button disappeared from the screen and playback of the question started. After the playback was over, a "Go" button appeared at the same location as the "Start" button together with the two visual stimuli, as shown in Fig. 4.1. Upon
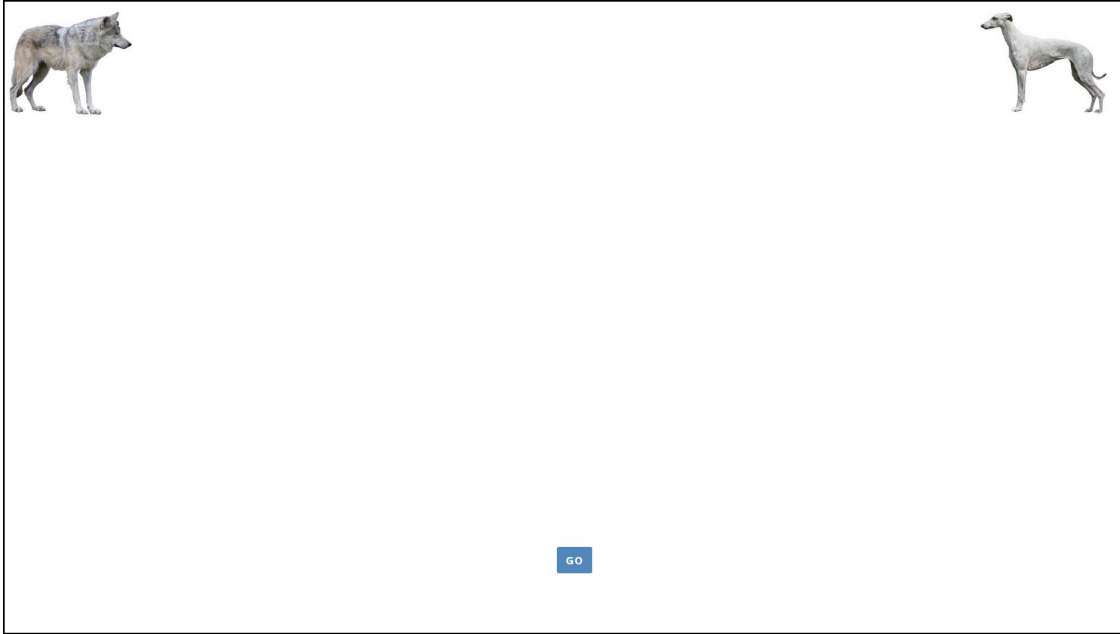
Figure 4.1 Sample visual world display. The display contains two pictures, one mentioned in the dialogue and an unmentioned competitor, each presented in one of the top corners of the screen.

clicking on the "Go" button, playback of the answer started together with the tracking of participants' mouse cursor from the "Go" button at the bottom of the screen towards one of the pictures at the top of the screen. Upon moving the cursor over one of the pictures, a choice was automatically made and participants were thus directed to the next trial.

Hypotheses

As discussed in the introduction, *eigentlich* and *tatsächlich* are particles which, in their modal usage, index how the utterance they are embedded in relates to a contextually-relevant propositional antecedent. Importantly, this entails that the presence of either one particle in a sentence is expected to affect incremental interpretation differentially depending on the specific modal meaning the particle signals in context. This general effect is expected to be modulated by the reliability of the input, as discussed above. While *eigentlich* signals a propositional contrast, *tatsächlich* signals a propositional agreement. Given the setup of the present experimental task, I am therefore interested in how the measured mouse trajectories

are affected by the presence of each particle in the test sentences. More concretely, I am interested in the shape of the trajectory segments measured after the onset of the particle and before the onset of the disambiguating noun, i.e., "Das ist PARTICLE [ein Bild von einem] NOUN". Crucially, any changes in mouse position within this time window can reasonably be assumed to be an effect of processing the particle, such that movements either towards or away from the image representing the referent mentioned in the question indicate a sensitivity to the meaning of the processed particle.

While the general account entertained here does not generate precise predictions about the shape of the mouse trajectories within the critical time window, it generates predictions about qualitative changes to the trajectories after the onset of the particles. For the **reliable** group, it predicts that upon hearing *eigentlich* participants will shift their mouse trajectory towards the referent not mentioned in the question, as that referent constitutes a contextually-relevant propositional alternative to the referent mentioned in the question. It also predicts that upon hearing *tatsächlich* participants will continue their trajectory towards the referent mentioned in the question. Note that, in both cases, if participants initiate any mouse movement prior to the onset of the particle, their trajectories are assumed to be en route towards the mentioned referent at the onset of the particle, as the mention of the referent in the question biases them to entertain that particular referent before any relevant semantic information from the response is processed.

For the **unreliable** group, the present account predicts the same qualitative patterns as in the **reliable** group but it also predicts an attenuation of those patterns, such that initiation of the mouse movements towards the selected referent might start later in the trial, or the curvature of the trajectories might be larger due to the higher degree of uncertainty.

Results

The data and analysis scripts are available for inspection under https://osf.io/dt8r3/.

Data cleaning and pre-processing

Prior to any analysis, the data set was cleaned according to both participant-based and response-based exclusion criteria. I first excluded any participant who selected a picture before the onset of the particle in more than 30% of their trials (n=3). This is motivated by the fact that any such decision would not be based on the processing of the relevant linguistic material in the test sentence. This exclusion resulted in a data set which included data from 47 participants.

After excluding three individual participants as per the criterion above, I excluded any individual trial in which a participant took longer than 1000 ms to make a decision after the offset of the disambiguating noun (5% of the data) as well as any trial with fewer than 10 observations (7% of the data), an observation corresponding to a single change in horizontal mouse position as sampled during the recording of the mouse movements. I also excluded all trials in which participants selected the competitor instead of the target referent, i.e., *eigentlich* trials where participants selected the referent mentioned in the question and *tatsächlich* trials in which participants selected the unmentioned referent. Finally, all **unreliable** trials were discarded, such that the remaining data only contained **reliable** trials.

The cleaned data, plotted and analyzed below, consists of mouse trajectories which were linearly extrapolated from the actual recorded trajectories. This procedure was adopted in order to circumvent the loss of relative temporal information when aggregating over trajectories with different absolute durations. Data points were extrapolated at every 10 ms time interval for each individual participant-item pair.

Confirmatory analyses

First I report and inspect the mean mouse trajectories, as shown in Figure 4.2.

The figure is split in two columns, each showing the data from one of the two experimental

144
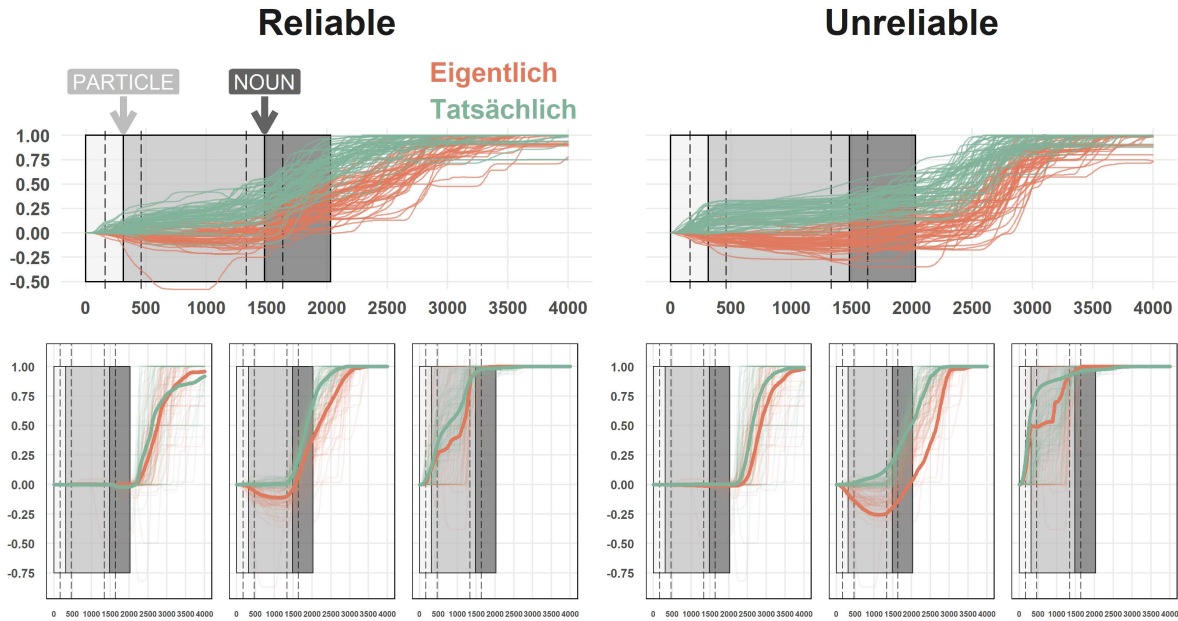
Figure 4.2 Mean mouse trajectories. **Top**: Mean item trajectories aggregated over participants. **Bottom**: Clusters of mean trajectories aggregated over items and participants, with mean item trajectories shown in the background. The shaded areas in the graphs show the (average) time windows between the onsets of the discourse particle and the disambiguating noun. The left column shows data from the **reliable** group, while the right column shows data from the **unreliable** group. *Tatsächlich* items are shown in green while *eigentlich* items are shown in orange.

groups, **reliable** on the left and **unreliable** on the right. Each graph at the top of the figure shows the individual item trajectories aggregated over participants. Trajectories unfold over time, mapped onto the x-axis, relative to the position of the target referent, mapped onto the top of y-axis. The shaded areas in the graphs show the (average) time windows between the onsets of the relevant linguistic material in the signal, namely the particle and the noun which disambiguates the referring expression. As the figure shows, there is considerable variation in terms of the shape of the mouse movements: while some trajectories show an early bias for the mentioned referent, either continuing en route towards it after the onset of the discourse particle or changing its course towards the unmentioned referent, others linger around the center of the screen until after the onset of the disambiguating noun. Given this variation in the data, instead of producing a single aggregate for each discourse particle I

first split the data into clusters of similarly shaped trajectories, as shown at the bottom of Fig. 4.2. The clustering of the data was done using an automated procedure implemented in the `mt_cluster()` function from the `mousetrap` package for R (Kieslich and Henninger, 2017). `mt_cluster()` generates a user-defined number of trajectory clusters by computing a matrix of distances between pairs of individual trajectories and then splitting it by the number of desired clusters. Before inspecting in more detail the aggregates shown within the clusters at the bottom of Fig. 4.2, which consist of mean trajectories aggregated over all experimental items and participants, I look at the relative proportion of trajectories per cluster, as shown in Figure 4.3.
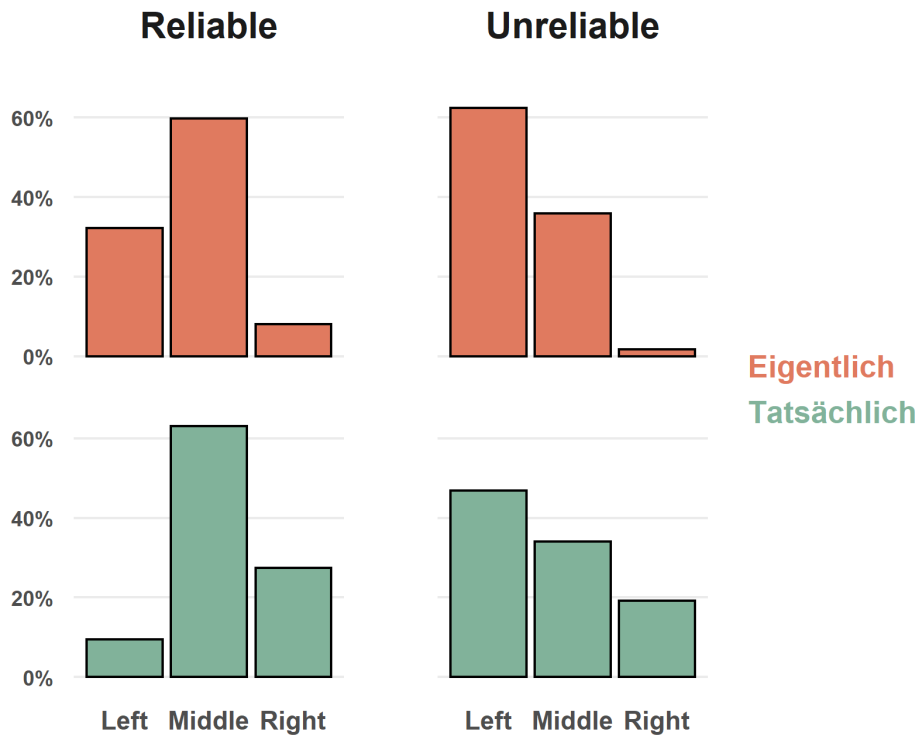
Figure 4.3 Proportion of trajectories per cluster. Proportion of data points per trajectory cluster in Figure 4.2. The left column shows data from the **reliable** group, while the right column shows data from the **unreliable** group. *Tatsächlich* items are shown in green while *eigentlich* items are shown in orange.

In rough terms, the clusters represent three types of trajectories: those in the left-hand cluster can be said to represent a strategy where one waits for disambiguation from the

noun; those in the middle cluster can be said to represent a hesitant interpretation of the particle; those in the right-hand cluster can be said to represent an early drift towards the referential target. As the figure shows, both in the **reliable** and **unreliable** groups the data is unevenly split across clusters. In the **reliable** group, 32% of the *eigentlich* trajectories are found in the cluster in the left-hand panel ($P(Y = left) = .32$, 95% CrI [0.19, 0.48]), 60% of the trajectories being found in the cluster in the middle panel ($P(Y = middle) = .60$, 95% CrI [0.42, 0.75]) and 8% in the cluster in the right-hand panel ($P(Y = right) = .08$, 95% CrI [0.03, 0.18]). Ten percent of the *tatsächlich* trajectories are found in the cluster in the left-hand panel ($P(Y = left) = .10$, 95% CrI [0.04, 0.19]), 63% of the trajectories being found in the cluster in the middle panel ($P(Y = middle) = .63$, 95% CrI [0.45, 0.78]) and 27% of them in the cluster in the right-hand panel ($P(Y = right) = .27$, 95% CrI [0.15, 0.44]). In the **unreliable** group, on the other hand, 62% of the *eigentlich* trajectories are found in the cluster in the left-hand panel ($P(Y = left) = .62$, 95% CrI [0.48, 0.76]), 36% of the trajectories being found in the cluster in the middle panel ($P(Y = middle) = .36$, 95% CrI [0.23, 0.50]) and only 2% in the cluster in the right-hand panel ($P(Y = right) = .02$, 95% CrI [0.00, 0.07]). Forty-seven percent of the *tatsächlich* trajectories are found in the cluster in the left-hand panel ($P(Y = left) = .47$, 95% CrI [0.32, 0.61]), 34% of the trajectories being found in the cluster in the middle panel ($P(Y = middle) = .34$, 95% CrI [0.19, 0.52]) and 19% of them in the cluster in the right-hand panel ($P(Y = right) = .19$, 95% CrI [0.12, 0.31]).

In order to quantify uncertainty over the results reported above, I fitted a Bayesian hierarchical model predicting cluster membership as a function of an interaction between the particles and the experimental groups, with *eigentlich* in the **reliable** group dummy-coded as the reference level. The model estimated the log odds of each observation being categorized as belonging to any given cluster, the estimates reported below being a conversion of (predicted) log odds to probabilities. It included the maximal random effect structure justified by the design, which in the present case is random slopes and intercepts for the

effect of particle for both participants and items. Since participants were assigned to only one experimental group, at the participant level I only estimate random slopes for the effect of group. The model, fitted using the R package `brms` (Bürkner, 2017) and described in detail in the scripts available in the supporting material, had the following form, shown in `brms` syntax:

$$\text{cluster} \sim \text{particle * group}$$
$$(1 + \text{particle} \mid \text{participant}) +$$
$$(1 + \text{particle * group} \mid \text{item}),$$
$$\text{family} = \text{categorical}()$$

I now turn back to Fig. 4.2 to inspect the actual trajectories within each of the clusters. As visual inspection of the mean trajectories within each cluster in the **reliable** group shows, within two of the three clusters, the trajectories for the two particles are qualitatively different from one another. The three distinct types of *tatsächlich* trajectories, shown in green, consist of shifts towards the target after the onset of the disambiguating noun (left panel, 10% of the data), shifts towards the target within the particle window (middle panel, 63% of the data), as well as shifts towards the target before the onset of the particle without subsequent change of course (right panel, 27% of the data). The *eigentlich* trajectories, shown in orange, consist of shifts towards the target after the onset of the disambiguating noun (left panel, 32% of the data), change-of-minds shifts towards the target within the particle window (middle panel, 60% of the data), as well as shifts towards the target before the onset of the particle without subsequent change of course (right panel, 8% of the data). Taken together, these results suggest that both *eigentlich* and *tatsächlich* are used predictively in the interpretation of the referring expressions, such that their presence in an utterance induces participants to move their mouse towards the referential target before hearing information from the disambiguating noun in 90% of *tatsächlich* trials and 68% of *eigentlich* trials.

Visual inspection of the mean trajectories within each cluster in the **unreliable** group

148

shows that, as in the **reliable** group, in two of the three clusters, the trajectories for *eigentlich* and *tatsächlich* differ qualitatively from one another. Crucially, while the qualitative patterns within each cluster are similar to those in the **reliable** group, the exact shape of the trajectories is different when participants are exposed to unreliable usages of the particles. Interestingly, in the case of the cluster in the left-hand panel, the shifts towards the target are delayed compared to the same cluster in the **reliable** group. In the cluster in the middle panel, the shift towards the target for *tatsächlich* items occurs earlier compared to the same cluster in the **reliable** group, following a more curved trajectory. The change-of-mind shift towards the target for *eigentlich* items, on the other hand, is exacerbated compared to the cluster in the **reliable** group, meaning that due to a larger initial bias for the mentioned referent the curvature of the trajectory is also more pronounced. Finally, in the cluster in the right-hand panel, the initial bias for the target for both *eigentlich* and *tatsächlich* items is exacerbated compared to the same cluster in the **reliable** group. Following this exacerbated initial bias, the trajectory for *tatsächlich* items is flatter than in the same cluster in the **reliable** group, while the trajectory for *eigentlich* items first flattens before shifting abruptly towards the target. Taken together, these results suggest that exposure to unreliable trials generally leads to temporal modulations of the qualitative patterns found in the **reliable** group, together with an overall reduction in the number of trials in which the particles are used predictively, from 90% to 67% in the case of *tatsächlich* and from 68% to 57% in the case of *eigentlich*. Moreover, in the case of the target bias found in the predictive trajectories, there is a modulation of the magnitude of the bias, with an increase in the bias in the **unreliable** group.

In order to better assess the impact the particles have on the resolution of reference, I focus on the shapes of the mouse trajectories within a critical time window spanning 150 ms after the onset of the particle up until 150 ms after the onset of the disambiguating noun, as shown at the top of Figure 4.4.

Crucially, any changes in the relative horizontal mouse position within this time window
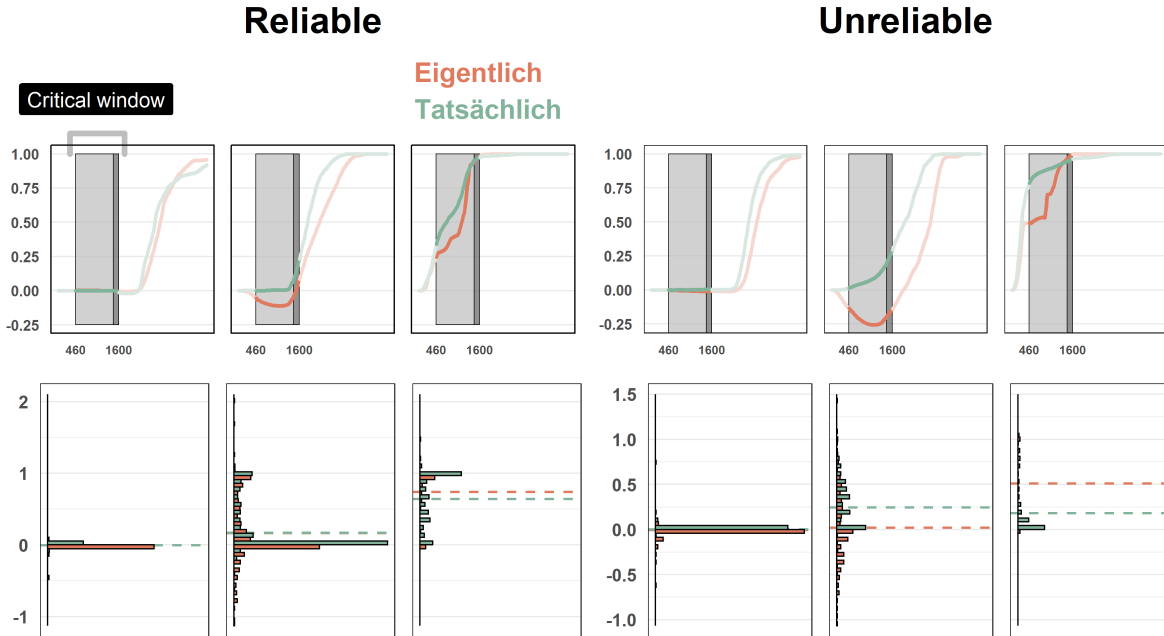
Figure 4.4 Horizontal mouse position within the critical time window. **Top**: Clusters of mean trajectories aggregated over items and participants, with the segments falling within the critical time window highlighted. The shaded areas in the graphs show the (average) critical time window, which starts at 150 ms after the onset of the particle and ends at 150 ms after the onset of the disambiguating noun. **Bottom**: Distributions of differences in the horizontal mouse position for each participant-item pair, shown in terms of each trajectory cluster. The dotted lines show the mean differences for each discourse particle. The left column shows data from the **reliable** group, while the right column shows data from the **unreliable** group. *Tatsächlich* items are shown in green while *eigentlich* items are shown in orange.

can reasonably be assumed to be an effect of processing the particle, such that movements either towards or away from the mentioned referent indicate a sensitivity to the meaning of the processed particle. In order to quantify such effects, I calculate for each participant-item pair the difference between the absolute horizontal mouse position at the start and end of the critical time window, which, on average, are at the 460 and 1600 ms marks, respectively. The computed differences are shown at the bottom of Fig. 4.4, with values around zero indicating no change in horizontal position, positive values indicating movements towards the target, and negative values indicating movements towards the competitor. Note that the target is different for each particle: while in the case of *tatsächlich* it is the referent

mentioned in the question, in the case of *eigentlich* it is the unmentioned referent. In order to quantify uncertainty over the results reported below, I fitted two Bayesian hierarchical models predicting differences in horizontal mouse position within each trajectory cluster as a function of an interaction between the particles and the experimental groups, with *eigentlich* in the **reliable** group dummy-coded as the reference level. Each model estimated changes in the magnitude of the horizontal position when making pairwise comparisons between clusters. Both models included the maximal random effect structure justified by the design, which in the present case is random slopes and intercepts for the effect of particle for both participants and items. Since participants were assigned to only one experimental group, only at the item level do I estimate random slopes for the effect of group. The models, fitted using the R package `brms` (Bürkner, 2017) and described in detail in the scripts available in the supporting material, had the following form, shown in `brms` syntax:

$$\text{horiz\_diff} \sim \text{particle} * \text{group}$$
$$(1 + \text{particle} \mid \text{participant}) +$$
$$(1 + \text{particle} * \text{group} \mid \text{item})$$

First, I report the mean differences within each cluster in the **reliable** group, anchoring them to the actual trajectories shapes. As visual inspection of the figure shows, in the cluster in the left-hand panel, both the *tatsächlich* and the *eigentlich* trajectories are flat, resulting in differences which are practically zero. In the cluster in the middle panel, the *tatsächlich* trajectory is flat through most of the critical window, suddenly shifting towards the target shortly before the onset of the noun, resulting in a difference of 0.17 ($\beta = 0.18$, 95% CrI [0.12, 0.25], $P(\beta > 0) = 1$). The *eigentlich* trajectory, on the other hand, shows a light curvature, first following an initial bias for the mentioned referent before changing course towards the unmentioned referent, resulting in a difference of 0.17 ($\beta = 0.16$, 95% CrI [0.08, 0.25], $P(\beta > 0) = 1$). In the cluster in the right-hand panel, the *tatsächlich* trajectory follows an initial bias for the mentioned referent, being mostly linear and resulting in a difference

of 0.64 ($\beta = 0.73$, 95% CrI [0.63, 0.83], $P(\beta > 0) = 1$). The *eigentlich* trajectory, on the other hand, follows an initial bias for the unmentioned referent before suddenly accelerating towards it midway in the critical window and thus resulting in a difference of 0.74 ($\beta = 0.77$, 95% CrI [0.62, 0.93], $P(\beta > 0) = 1$).

As for the cluster in the left-hand panel in the **unreliable** group, visual inspection of the figure shows that, much like in the same cluster in the **reliable** group, both the *tatsächlich* and *eigentlich* trajectories are flat, resulting in differences which are practically zero. In the cluster in the middle panel, the *tatsächlich* trajectory follows an initial bias for the mentioned referent, being more curved than the same trajectory in the **reliable** group and thus resulting in a difference of 0.25 ($\beta = 0.30$, 95% CrI [0.20, 0.41], $P(\beta > 0) = 1$). The *eigentlich* trajectory, on the other hand, shows a strong curvature, first following an initial bias for the mentioned referent before changing course towards the unmentioned referent, resulting in a difference of practically zero. In the cluster in the right-hand panel, the *tatsächlich* trajectory follows an initial bias for the mentioned referent, being more curved than the same trajectory in the **reliable** group and thus resulting in a difference of 0.18 ($\beta = 0.31$, 95% CrI [0.15, 0.47], $P(\beta > 0) = 1$). The *eigentlich* trajectory, on the other hand, starts mostly flat before suddenly accelerating towards the unmentioned referent midway in the critical window, thus resulting in a difference of 0.51 ($\beta = 0.65$, 95% CrI [0.35, 0.95], $P(\beta > 0) = 1$).

4.3   Discussion

In the present study, I investigated the role modal-marking discourse particles play in incremental interpretation and the online resolution of reference. To that end, I used a mouse-tracking task where participants were asked to listen to mini dialogues which included referring expressions containing the German particles *eigentlich* and *tatsächlich*. For each dialogue, participants were presented with two images on a screen and instructed to select

the one that best portrayed the topic of the exchange. Crucially, I was interested in the potential differential effects of processing *eigentlich* and *tatsächlich* in context, as measured in terms of any relative changes to the position of the mouse after hearing the particle. I was also interested in the effects of varying the pragmatic reliability of the particles, as manipulated in a separate condition which included both reliable and unreliable particle usages.

The results show that, in the reliable group, both *eigentlich* and *tatsächlich* were used predictively in the interpretation of the referring expressions, such that their presence in an utterance induced participants to move their mouse towards the relevant referential target – which was different for each particle – before hearing information from the disambiguating noun. The results also show that exposure to unreliable particle usages not only led to temporal modulations of the qualitative patterns found in the reliable group but it also reduced the predictive potential of the particles, as evidenced by an overall reduction in the number of predictive trajectories.

What these results suggest is that, as found for the Dutch particles *eigenlijk* and *inderdaad*, discourse particles with a modal-marking function do affect incremental interpretation, at least in the context of reference resolution. Crucially, they affect not only the outcome of the interpretive process but also how it unfolds in real-time. While the contrast-marking particle *eigentlich* modulates interpretation insofar as it induces comprehenders to revise an assumption at issue in the discourse, e.g., that the discourse referent under consideration is actually a dog as opposed to a wolf, the agreement-marking particle *tatsächlich* modulates interpretation insofar as it induces comprehenders to confirm a prior assumption at issue in the discourse, e.g., that the discourse referent under consideration is indeed a dog as opposed to a relevant alternative like a wolf.

These findings raise a number of questions about the status of modal particles both in relation to discourse particles more generally as well as in relation to other linguistic markers of epistemicity. Regarding their role as particles with the specialized function of marking

modality, the present findings illustrate how modal particles can impact the interpretation of their host utterances, underlining their modal character and indeed warranting their special status among the more general class of discourse particles. At the same time, while in the present study the materials were designed so as to explicitly elicit a modal interpretation of *eigentlich* and *tatsächlich*, in van Bergen and Bosker (2018), for instance, the materials allowed both a modal as well as a non-modal interpretation of *eigenlijk* and *inderdaad*. As their results regarding *eigenlijk* suggest, a particle's modal meaning might not necessarily be the most salient meaning in context, at least not when the two meanings are relatively similar to one another, as in the case of *eigenlijk*/ *eigentlich* and *inderdaad*/ *tatsächlich*, which signal complex pragmatic meanings which are hard to distinguish from one another.

As discussed before, however, the difference between marking a general sense of (dis)agreement between two utterances and (dis)agreement between two propositions might be one of degree as opposed to type. Indeed, even in the case of particles which signal arguably quite different meanings depending on their particular usage – such as *bloß* and *nur*, which index either a subjective sense of strengthening or a restriction depending on what they relate their host utterance to – the difference between a particle's modalized and non-modalized meaning might not be as apparent as it seems. For instance, Dörre and Trotzke (2019) argue that there is a "common semantic denominator" between the two senses of *bloß* and *nur*, one which involves restriction of a particular relevance domain. In the case of, *eigenlijk*/ *eigentlich* and *inderdaad*/ *tatsächlich*, both meanings are related to the management of expectations in discourse, whether those are expectations to do with concrete propositions which can be modalized over or expectations related to the flow of conversation and the general maintenance of understanding, which can nevertheless also be signaled upon.

Regarding their role as markers of epistemicity, modal particles should be considered vis-à-vis other such markers, like other lexical modal markers as well as morphosyntactic and prosodic modulations of the linguistic signal, to highlight only canonical linguistic means of marking epistemicity. Indeed, prosodic as well as morphosyntactic cues can be used to

mark one's stance in discourse, both in the absence of and in conjunction with particles. For instance, let's look back at examples (4)-(6) from the introduction, reproduced here as (18)-(20).

A: Did you see how fast that bird flew by?

(18) B: *Really*? I didn't even notice.

(19) B: I don't *really* think it flew by that fast.

(20) B: It *really* did fly by fast.

In the examples above, B uses many different strategies in addition to their usage of *really* to mark their stance with regard to the content of A's utterance: in (18), B employs the particle *even* to emphasize an aspect of their own response to A's statement; in (19), B employs the intensifier *that* to highlight information about the degree of the event at issue; in (20), B employs a non-canonical syntactic structure to emphasize the validity of A's statement; finally, both in (19) and (20) a pitch accent could be used to signal different degrees of epistemic belief in the strength of the respective propositions, i.e., "I don't REALLY think it flew by that fast" vs. "I don't really think it flew by THAT fast" (19) and "It REALLY did fly by fast" vs. "It really did a bird" in (20). This illustrates not only the variety of different linguistic strategies which can be used to mark epistemicity in the case of English but also the expressive potential of combining different strategies.

As shown by Kurumada et al. (2014), for instance, combining different cues can help guide comprehension in context, as when prosody modulates the interpretation of a lexical expression like "it looks like". Similarly, as shown by Dörre and Trotzke (2019), a non-lexical marker can be used to constrain the meaning of a lexical marker even in the absence of otherwise overt contextual cues, as when pitch accent disambiguates the meaning of a modal particle. All in all, what this suggests is that modal particles should be seen as yet another tool in the toolkit of modality markers, one which shows functional behaviors similar to those of other modal tools despite, from a formal perspective, belonging to a special

class of markers. In that vein, future research could address the role of competing modal cues in incremental interpretation, such as when pitch accent and a modal particle provide conflicting information in context. Is interpretation affected by such a pragmatic mismatch between different linguistic sources of information? Does one cue take precedence over the other, such that maybe specialized markers like particles are interpreted more or less reliably depending on whether non-lexical cues like pitch accent or word order alternation are used unreliably? The findings of the present study provide evidence that a pragmatic mismatch in the form of distributional unreliability in the usage of the particle itself does affect the predictive protential of a particle in context. Future research should investigate the role played by other forms of pragmatic mismatch in the interpretation of modal particles and lexical markers of stance more generally.

In terms of the cognitive nature of the meanings associated with a polyfunctional particle, one could ask whether their modal meaning is represented differently than their non-modal counterpart. Since in the case of some particles the two meanings seem to be intertwined not only diachronically but also synchronically, with subtle differences in terms of their exact synchronic functions, that raises questions about the nature of not only the representations themselves but also the interpretive processes underlying the understanding of modal particles in context. While earlier work suggested that the more conventionalized the modal meaning of a particle, the harder it is to process it (Dörre, Czypionka, et al., 2018), the present findings suggest a role of pragmatics in guiding the interpretation of particles in context. Future research could thus more directly address these sorts of questions, dissecting the exact dynamics of modal meaning representation.

All in all, the findings of the present study provide evidence that the particles *eigentlich* and *tatsächlich* serve as predictive cues in the incremental processing of language, allowing comprehenders to anticipate lexical material in a sentence before actually processing that material. These particles thus seem to function as pragmatic cues to interpretation, modulating the illocutionary force of the utterance they occur in. However, the nature of the

modulation is one which depends on the modal meaning associated with either particle, suggesting an interpretive process akin to the derivation of conventional implicatures, that is, one which builds on both conventionalized (non-truth-conditional) meaning as well as pragmatic expectations.

REFERENCES

Abraham, Werner (2016). "Discourse marker = discourse particle = thetical = modal particle? A futile comparison". In: *Discourse Particles: Formal Approaches to their Syntax and Semantics*. Ed. by Josef Bayer and Volker Struckmeier. De Gruyter, pp. 241–280. DOI: doi:10.1515/9783110497151-010. URL: https://doi.org/10.1515/9783110497151-010.

Bürkner, Paul-Christian (2017). "brms: An R package for Bayesian multilevel models using Stan". In: *Journal of statistical software* 80.1, pp. 1–28.

Döring, Sophia and Sophie Repp (2019). "The modal particles ja and doch and their interaction with discourse structure: Corpus and experimental evidence". In: *Experiments in Focus*.

Dörre, Laura, Anna Czypionka, et al. (2018). "The processing of German modal particles and their counterparts". In: *Linguistische Berichte* 255, pp. 58–91.

Dörre, Laura and Andreas Trotzke (2019). "Chapter 6 The Processing of Secondary Meaning: An Experimental Comparison of Focus and Modal Particles in Wh-questions". In: *Secondary Content: The Semantics and Pragmatics of Side Issues*. Leiden, The Netherlands: Brill, pp. 143–167. ISBN: 9789004393127. DOI: https://doi.org/10.1163/9789004393127_007. URL: https://brill.com/view/book/9789004393127/BP000006.xml.

Fedriani, Chiara and Andrea Sansó (2017). "Pragmatic Markers, Discourse Markers and Modal Particles. What do we know and where do we go from here?" In:

Feng, Guangwu (2011). "A neo-Gricean pragmatic analysis of Chinese pragmatic markers". In: *Language Sciences* 33.3, pp. 417–434. DOI: https://doi.org/10.1016/j.langsci.2010.10.005.

— (2010). *A Theory of Conventional Implicature and Pragmatic Markers in Chinese*. Leiden, The Netherlands: Brill. DOI: https://doi.org/10.1163/9789004253179.

Fischer, Kerstin (2006). "Towards an understanding of the spectrum of approaches to discourse particles: introduction to the volume". In: *Approaches to discourse particles*. Brill, pp. 1–20.

Gast, Volker (2008). "Modal particles and context updating – the functions of German ja , doch , wohl and etwa 1". In:

Heine, Bernd et al. (2021). *The Rise of Discourse Markers.* Cambridge University Press. DOI: 10.1017/9781108982856.

Jucker, Andreas H. and Yael Ziv (1998). "Discourse markers : Introduction". In: *Pragmatics and beyond. New series* 57, pp. 1–12.

Kieslich, Pascal J. and Felix Henninger (2017). "Mousetrap: An integrated, open-source mouse-tracking package". In: *Behavior Research Methods* 49, pp. 1652–1667.

Kurumada, Chigusa et al. (2014). "Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings". In: *Cognition* 133.2, pp. 335–342.

Maschler, Yael and Deborah Schiffrin (2015). "Discourse markers: Language, meaning, and context". In: *The handbook of discourse analysis* 1, pp. 189–221.

Rasenberg, Marlou, Joost Rommers, and Geertje van Bergen (2019). "Anticipating predictability: an ERP investigation of expectation-managing discourse markers in dialogue comprehension". In: *Language, Cognition and Neuroscience* 35, pp. 1–16.

Roettger, Timo B and Michael Franke (2019). "Evidential strength of intonational cues and rational adaptation to (un-) reliable intonation". In: *Cognitive science* 43.7, e12745.

Schoonjans, Steven (2013). "Modal particles: Problems in defining a category". In: *Discourse markers and modal particles: Categorization and description*, pp. 133–161.

Unger, Christoph (2016). "DEGREES OF PROCEDURE ACTIVATION AND THE GERMAN MODAL PARTICLES JA AND DOCH - PART 3". In: *Studia Linguistica Universitatis Iagellonicae Cracoviensis* 2016, pp. 31–45.

Bergen, Geertje van and Hans Rutger Bosker (2018). "Linguistic expectation management in online discourse processing: An investigation of Dutch inderdaad 'indeed' and eigenlijk 'actually'". In: *Journal of Memory and Language*.

Chapter Five

Interpreting utterances with adjectival markers of epistemic stance and tracking non-literality in online language processing

5.1 Introduction

In Chapter 4, I presented a study focused on how comprehenders process utterances containing specialized lexical markers of common ground. More specifically, I investigated whether comprehenders draw predictive inferences about upcoming referential information when processing the German discourse particles *eigentlich* and *tatsächlich*, which mark relations of contrast and agreement between their host utterance and propositional information previously established in the discourse. Crucial for my analysis was the fact that linguistic markers such as *eigentlich* and *tatsächlich* encode in their lexical meaning the relevant relations – or procedures – to be established in context, despite being polyfunctional devices which mark different types of relations depending on the particular discourse context in which they are used. In the present chapter, I focus on a different class of lexical epistemicity markers,

namely pragmatic markers of stance.

In contrast to the German adverbs investigated in Chapter 4, here I investigate English adjectives which, crucially, do not serve a specialized function as discourse particles but rather have a conventionalized pragmatic meaning as intensifiers, i.e., as markers of subjective epistemic stance, alongside their lexical semantic meaning. In concrete terms, I'm interested in understanding whether intensifying adjectives, as pragmatic markers of stance, affect incremental interpretation in online processing, as measured in a word-by-word self-paced reading task. Crucially, the particular question I'm interested in is whether the English adjectives *actual, literal, real* can be used as predictive cues in the incremental processing of nominal metaphors, such that, in combination with a noun used metaphorically (e.g. *He is a weasel*), they cue a non-literal interpretation either more or less strongly than non-intensifying adjectives whose lexical semantics are either more or less compatible with a metaphoric interpretation (e.g. *He is a devious weasel* vs. *He is a furry weasel*). I find that intensifying adjectives cue a non-literal interpretation as strongly as non-intensifying adjectives which are interpreted more metaphorically (e.g. *He is a devious weasel*) while at the same time cueing a non-literal interpretation more strongly than non-intensifying adjectives which are interpreted less metaphorically (e.g. *He is a furry weasel*). Before introducing the experiment and reporting my findings in detail, I start by reviewing previous work on the function and interpretation of stance markers, with a particular focus on lexical markers of intensification like *really* in *He is really tall* or *absolute* in *It was an absolute mess*.

5.1.1 Stance-marking in language

As discussed in the previous chapter, language users have at their disposal in the languages of the world a variety of so-called epistemcity markers, linguistic devices which qualify the content of utterances with regard to how true or likely the communicated state of affairs

might be. While one common function of such devices is to mark the source of information a language user has access to, and how that in turn constrains what they might be trying to communicate, another fundamental function associated with epistemicity expressions is to mark a language user's personal commitment to the truth, plausibility or even relevance of what they are saying. Although, traditionally, these two functions have been associated with separate, relatively fixed classes of expressions (see Boye, 2012; de Haan, 2001) – known as evidentials and epistemic modals, respectively – more recent work on the expression of modality has acknowledged that, empirically, the boundaries between the two are much fuzzier (see work cited in the next paragraph), such that the same linguistic expressions can have evidential, epistemic, as well as possibly even other modal (and non-modal) meanings, often in the very same context of utterance.

Zooming in on the function traditionally associated with the class of epistemic modals, that is, to qualify or evaluate a particular piece of information uttered in discourse (Cornillie and Pietrandrea, 2012), epistemicity markers can be understood as markers of 'stance', a language user's personal attitude or assessment. Linguistically, stance can be expressed via grammatical as well as lexical means, the precise boundaries and relationship between one and the other being the source of much debate in the literature (Cornillie, 2009; Ricci and Rossari, 2018), both when analyzing specific stance markers within individual languages as well as when delineating their cross-linguistic status. Importantly, it has been argued that not only the linguistic status of the devices used in an assessment can vary on a continuum from lexical to grammatical but also the meaning of the assessment proper can vary between what is prototypically considered to be the marking of personal feeling/ attitude and the marking of the status of knowledge (Gray and Biber, 2012).

On the grammatical end of the continuum, stance is known to be marked via morphology, such as in languages containing evidential or mirative affixes (Aikhenvald, 2004, 2012; de Haan, 2012) or evaluative/ expressive morphology (Grandi and Körtvélyessy, n.d.; Körtvélyessy, 2015). While morphological markers of evidentiality and mirativity tend to be of

162

obligatory usage, evaluative morphemes are optional, much like their lexical counterparts in languages without grammaticalized means of marking stance. Indeed, on the lexical end of the continuum, stance is commonly expressed via modal or cognition verbs (e.g. 'might' and 'think' in English, respectively) and adverbs (e.g. 'really' in English), as well as via particles of different sorts, such as the ones tackled in Chapter 4. Among the specific functions lexical markers of stance can serve in discourse is the marking of the degree to which a language user commits to the truth or validity of the information they are conveying in their utterance – say, for instance, that someone is *really* tall, or that a particular situation was an *absolute* mess. In such constructions, *really* and *absolute* indicate that the person was tall and that the situation was a mess to an extent which is above the usual threshold of tallness and messiness, even if these thresholds themselves are subjective and dynamic such that they change according to the situation and time of utterance. What's important is that, in their function as modifiers of degree, stance markers are used to intensify the relevant semantic property associated with the modified item, as the adverb *really* does when it modifies the adjective *tall*, and as the adjective *absolute* does when it modifies the noun *mess*. Crucially, such modifiers impart a subjective assessment to the propositional meaning they intensify, speaking to their role as markers of stance.

In English, adverbs and adjectives act as prototypical intensifiers, usually modifying adjectives (e.g. *He is really tall*) and nouns (e.g. *It was an absolute mess*), respectively. Even otherwise functionally very similar intensifiers like *really* and *absolute* can have subtle differences in interpretation depending not only on the particular discourse and linguistic context in which they occur but also on their nature as lexical items in and of themselves, such that their lexical semantics might become more or less apparent in different modificational contexts (Paradis, 2012). For instance, while *really* in *He is really tall* seems to rather straightforwardly intensify *tall*, without clear traces of any epistemic meaning related to 'reality' (in the sense of literal truth), *absolute* in *It was an absolute mess* might be said not only to indicate an intensified degree of *mess* but also to impart some sense of 'absoluteness'

163

to it (in the sense of completion), such that the adjective's conventionalized meaning of intensification is not clearly distinguishable from its lexical semantic meaning.

Previous work on intensification has established that different lexical items are more or less prone to imparting their lexical semantics onto their meaning as markers of stance (Athanasiadou, 2007; Ghesquierè, 2017). Indeed, while some degree modifiers in English like *very* and *quite* have lost any lexical meaning of their, only serving as modifiers of degree in a strict grammatical sense, other markers like *absolute(ly)* and *real(ly)* maintain, at least in some usages, their lexical meaning as adjectives and adverbs. In the case of adverbial markers like *really*, the different meaning profiles correlate with particular syntactic configurations, such that the farther the adverb is from the adjective in a sentence the more of an objective, and thus lexical, meaning it tends to have. For example, in *He really is tall* the adverb signals a subjective emphasis on the truth of the proposition at hand, which relates to its meaning of 'reality', while in *He is really tall* the same adverb signals a reinforcement of the scalar degree of *tall*, which only indirectly relates to any objective truth of such property. In the case of adjectival markers like *real* and *absolute*, on the other hand, their meaning as markers of intensification is not always clearly distinguishable from their lexical meaning, especially as, in English, they mostly occur right before the modified noun (e.g. *It was an absolute mess*), which means that their syntactic position alone provides no reliable cue to their meaning.

One explanation as for why some expressions have an epistemic meaning profile while others don't might be that intensifiers which have evolved from truth markers retain part of their truth-endorsing meaning (Bardenstein and Ariel, 2021; Heine et al., 2002): for example, while *really* functions both as an intensifier (*He is really tall*) and as an emphasizer (*He really is tall*), partially preserving its original meaning related to truth in the latter usage, in the same construction *absolutely* seems to function only as an intensifier (*He is absolutely tall/ #He absolutely is tall*), unless it is used in a more parenthetical sense. Other than their syntactic position, prosodic focus can serve as an additional cue as to

what the exact meaning contribution of a lexical marker might be: while both *really* and *absolutely* can receive prosodic stress in their intensificational usage (*He is REALLY tall/ He is ABSOLUTELY tall*), a pitch accent on the verb seems to co-occur more naturally with *really* than with *absolutely* (*He really IS tall/ #He absolutely IS tall*), whereby *really* is more flexible in that it can both emphasize the validity of the predicate as a whole and the scalar degree of *tall* in particular.

In the present study, I'm interested in understanding how lexical markers of intensification, as discussed so far, affect *incremental* interpretation in online processing, as measured in a word-by-word self-paced reading task. Crucially, the particular question I'm interested in is whether the English intensifying adjectives *actual*, *literal*, and *real* can be used as pragmatic cues in the incremental processing of nominal metaphors, such that they cue a likely non-literal interpretation in spite of their lexical semantic meaning relating to truth. Before introducing the experiment and reporting my findings, I review previous findings on the processing of adjectival modification more generally.

5.1.2   Processing profile of adjectival modification

Adjectives are known to act as predictive cues to interpretation in the online processing of language. Classic findings from the psycholinguistic literature show that comprehenders are able to modulate their incremental interpretation of a linguistic signal after processing a prenominal adjective, such that they are able to predict a referent even before hearing the relevant noun in a referring expression (Sedivy et al., 1999), namely by drawing a contrastive inference between potential referents in context on the basis of information from the adjective. Such early findings have been replicated multiple times in more recent studies (Gardner et al., 2021; Grodner and Sedivy, 2011), an additional finding being that comprehenders are able to suspend any predictive inferencing when finding evidence in the input that the utterance producer is unreliable.

165

Further results suggest that the processing of adjective-modified nouns phrases is highly efficient regardless of where the modifier occurs relative to the noun (Rubio-Fernandez and Jara-Ettinger, 2020), as in languages with postnominal adjectival modification comprehenders draw predictive inferences based not on properties (e.g. blueness), the prototypical meaning of an adjective in a referring expression, but rather on kinds (e.g. triangles), information which is usually derived from nouns rather than adjectives. Additionally, adjectives seem to serve as predictive cues in processing even in cases of referential overspecification (Tourtouri et al., 2021; Wu and Ma, 2020), the type of adjective being crucial as to whether it is processed predictively or not.

Across these different studies, the sorts of adjectives that have been investigated included color adjectives (e.g. *blue*), material adjectives (e.g. *glass*), as well as vague scalar adjectives (e.g. *tall*). The key finding from this body of work is that, given general discourse expectations, comprehenders can use various types of adjectives as predictive cues in incremental interpretation, being able to predict referents before processing disambiguating semantic information from the nouns modified by the adjectives. Crucially, in these studies the adjectives provided cues to the literal meaning of the referring expressions, usually descriptions of objects of different colors and shapes.

In the present study, my goal was to test whether comprehenders process non-literal, metaphoric statements predictively depending on the type of adjective used to modify the noun in the predicate of the metaphoric construction (e.g. *He is a(n) ADJ weasel*). Crucially, unlike previous psycholinguistic work focused on the processing of adjective-modified noun phrases, here I'm interested in how comprehenders process expressions containing adjectives which function as pragmatic markers of intensification despite having a lexical semantics related to truth or literality. What this means is that, unlike in the case of expressions modified by adjectives with a truth-conditional semantic meaning, expressions modified by adjectives with a pragmatic meaning of intensification might be temporarily ambiguous not because they are denotationally underspecified but rather because the adjectives' pragmatic

and semantic meanings might induce different, if not conflicting, interpretations during online processing.

Based on the work discussed in the previous section, which highlights the non-trivial interplay between the pragmatic, stance-marking meaning and the lexical semantic meaning of both adverbial and adjectival intensifiers, intensifying adjectives are expected to affect incremental interpretation differentially depending on whether they cue an interpretation more or less congruent with the most likely interpretation in context, whatever that may be. Crucially, in a metaphoric statement like *He is a real weasel sometimes during games*, where *weasel* refers to a person who is sneaky or devious, the adjective *real* can be said to signal both a non-literal (i.e. pragmatic) as well as a literal (i.e. semantic) meaning. At face value, this might suggest a potential (temporary) ambiguity in interpreting the utterance, at least until a cue like *sometimes*, which signals a likely non-literal interpretation of *He is a real weasel*, enters the frame and further constrains the interpretation process. However, this *prima facie* ambiguity is not expected under the assumption that comprehension involves dynamically weighing the cues available in a signal at any given moment in processing and deriving therefrom the most likely interpretation given general discourse expectations (Degen and Tanenhaus, 2019; Rohde and Kurumada, 2018), including the expectation that language is produced in informative ways given the producer's discourse goals. Indeed, under such an expectation-based account of (incremental) interpretation, the pragmatic meaning of intensification of *real* in the case above is arguably more congruent with a non-literal interpretation of *weasel* than its semantic meaning of 'reality', which therefore suggests no qualitative ambiguity in interpreting the utterance, assuming that the former takes precedence over the latter. Theories which posit that cue weighing is based on some form of expectancy mechanism need to be grounded in empirical measures of processing via an explicit linking function. In the present study, I assume that the updating of discourse expectations is linearly correlated with reading times measured at individual words in a sentence, such that the less a word is expected in a particular discourse context the more

167

difficult it is to process and integrate.

In the section below, I describe the exact predictions of the present study[1], as they relate to the test items, as well as the details and setup of the experimental task. The reported study was pre-registered under https://osf.io/vdjc4 and, unless otherwise noted, all analytic choices and procedures follow the pre-registered analysis plan.

5.2 Experiment

Method

Participants

80 self-reported monolingual speakers of English were recruited using the crowdsourcing platform Prolific. Participants received £2.18 for their participation in the experiment, which lasted on average no longer than 15 minutes. Data collection was conducted entirely online and all collected data was stored at servers from the Osnabrück University in Germany.

Materials and design

Participants were asked to read 33 English sentences. Their task was to read each sentence for comprehension, according to the procedure described below. There were 15 critical trials containing adjective-modified metaphors, as in the examples 1-3 below, as well as 18 filler trials, 12 of which also contained adjective-modified metaphors (e.g. *He is the liveliest fish in the tropical aquarium*) and 6 of which contained literal, non-metaphoric statements (e.g. *This afternoon we should take a drive in the countryside*).

---

[1]This study has been conceived in collaboration with Hannah Rohde, based on the original work reported in Naylon-Roberts (2016), an unpublished bachelor's thesis supervised by Hannah Rohde at the University of Edinburgh. While the design of the present experiment follows, with the author's consent, the design of the original experiment reported in Naylon-Roberts (2016), all experimental items and data reported here are novel, meaning that any analyses and interpretation that follow are also novel and entirely independent.

All critical statements consisted of nominal metaphors, such that the topic always consisted of a third-person singular pronoun (either *he* or *she*) and the vehicle, i.e., the noun in the predicate of the construction, always consisted of an animal conventionally associated with a psychological or physical attribute characteristic of people (e.g. *weasel* ~ a person who is sneaky or devious). The critical sentences always contained the adverb 'sometimes' after the noun, the adverb functioning as a strong cue to non-literality. There were 15 critical animal nouns (see the appendix for the item list), meaning that, on each trial, participants were presented with an altogether different metaphor. Crucially, each critical item appeared in one of three experimental conditions, as in the triplet below:

Metaphor-supporting He is a *devious* weasel sometimes during games.

Stance-marking He is an *real* weasel sometimes during games.

Literal-supporting He is a *furry* weasel sometimes during games.

The only difference between the sentences above are the adjectives which modify the vehicle in the metaphor. Crucially, by design, *devious weasel* is meant to cue an interpretation which is more likely metaphoric, i.e., one where a sneaky or devious person is being referred to; *furry weasel*, on the other hand, is meant to cue an interpretation which is more likely literal, i.e., one where what is being referred to is the actual animal and not a person. As discussed in the introduction, the adjective *real* in such a construction functions as a stance marker, i.e., as a marker of the attitude or subjective stance of the utterance producer, more specifically one which can be construed as intensifying the meaning of *He is a weasel*. Crucially, aside from an adjective-modified noun phrase, all critical items also contained a modifier adverbial phrase headed by *sometimes*. By design, *sometimes* was meant to cue a likely metaphoric interpretation of the sentence, which, from an incremental perspective, either supports or clashes with the unfolding interpretation, depending on whether that interpretation is more likely metaphoric, as should be the case in the *Metaphor-supporting* items

169

(i.e. *He is a devious weasel ...*), or literal, as should be the case in the *Literal-supporting* items (i.e. *He is a furry weasel ...*).

Procedure

Written instructions were provided prior to the actual task, followed by three practice trials which mimicked the exact procedure of the test trials. On each trial, participants were presented with a target sentence, as described in the section above. The sentence was masked at the beginning of the trial, underscores marking the location of each individual word. Above the underscores was the written cue "Press the SPACE bar to reveal the words". As indicated by the cue, participants were instructed to press the space bar in order to reveal each word in the sentence. Crucially, this means that they read the sentence one word at a time, in a self-paced manner. Once they read the last region in the sentence, pressing the space key would trigger a button labeled "Next", which had to be clicked in order to advance to the next trial. Once every third trial, instead of the "Next" button, participants were presented with a comprehension question relating to the sentence they read. They were asked to choose between two response options and received written feedback regarding their choice, which had the sole purpose of keeping participants engaged with the task of reading for comprehension.

Hypotheses

The hypothesis entertained here is that the type of adjective found in the items should affect incremental sentence interpretation differentially depending on whether the semantics of the adjective cues an interpretation more or less congruent with a non-literal, metaphoric meaning. Importantly, all predictions alike relate to the processing signatures measured at the adverb *sometimes*, which is the region directly following the modified noun in the critical sentence frame. This entails an effect of adjective type such that sentences with a literal-supporting adjective (*He is a furry weasel sometimes ...*) are expected to be harder

170

to process at the adverb *sometimes* compared to sentences with a metaphor-supporting adjective (*He is a devious weasel sometimes ...*), given that the adverb provides a cue which conflicts with a literal interpretation of *He is a weasel.* Crucially, sentences with a stance-marking adjective (*He is a real weasel sometimes ...*) are expected to be processed more easily at the adverb compared to sentences with a literal-supporting adjective, however, only if the stance-marking adjectives are taken as pragmatic cues to non-literal meaning, i.e., if they are interpreted not in terms of their semantics related to truthfulness or reality but rather in terms of their conventionalized pragmatic meaning as intensifiers. If, on the other hand, they are interpreted in terms of their truth semantics, then sentences with a stance-marking adjective are expected to be processed similarly to sentences with a literal-supporting adjective.

Note that these predictions are qualitative in nature: they relate to the overall *directionality* of a given effect, and not to the precise quantitative relations between the empirical response distributions of any two experimental conditions.

Results

The data and analysis scripts are available for inspection under https://osf.io/5v9xf/.

Data cleaning

Prior to any analysis, the data set was cleaned according to criteria defined in the pre-registered analysis plan, which include both participant-based and response-based exclusion criteria. I first excluded the data from any participant who self-reported being a monolingual speaker of a language other than English (n=1). I then further excluded any participant who did not reach an accuracy of 80% on the comprehension questions (n=0). Finally, I excluded any participant who deviated in more than 30% of their trials in more than 2.5 standard deviations from the grand mean for a given condition, as measured at any individual sentence region (n=0). Altogether, these exclusions resulted in a data set which included data from

79 participants.

In addition to excluding one participant as per the criteria above, I excluded any individual data point which deviated in more than 2.5 standard deviations from the grand mean for a given condition (1% of the data). Finally, I removed any data point lying outside of a pre-established range of interest set between 100 and 1000 ms (2% of the data). This reflects a meaningful range for the reading of individual words in self-paced reading (though see Jegerski and vanPatten (2013) on setting absolute cut-off points).

Confirmatory analyses

Before turning to the main results, which are based on aggregates over all experimental items and participants, I report and inspect the general reading time distributions. Figure 5.1 shows the empirical response distributions[2] for each condition at each sentence region.

As the figure shows, there is considerable overlap between the distributions at all sentence regions, which is to be expected as, even in the case of reliable effects, the usual mean word-level reading time differences in semantic and pragmatic processing are in magnitudes smaller than 50 ms. At the adjective (ADJ), noun (weasel), and adverb (sometimes) regions, where the crucial semantic information is read, the distributions show heavier right tails compared to other sentence regions, suggesting that these words are harder to parse and integrate, which is to be expected given their informational load in the comprehension task. Let's now turn to Figure 5.2 to inspect the grand means .

The figure shows that already at the adjective items in the *Metaphor-supporting* condition are read descriptively faster than items in the *Literal-supporting* condition, suggesting that even before information from the metaphoric noun enters the frame interpretation may

---

[2]In the pre-registered analysis plan, I committed to normalizing the reading times at each word by their number of syllables, given a potential analysis of the RTs at the adjectives and nouns, which vary across trials. However, given that the confirmatory analysis is focused only on the adverb region, which contains the same word across all trials, I ultimately decided not to normalize the reading times.
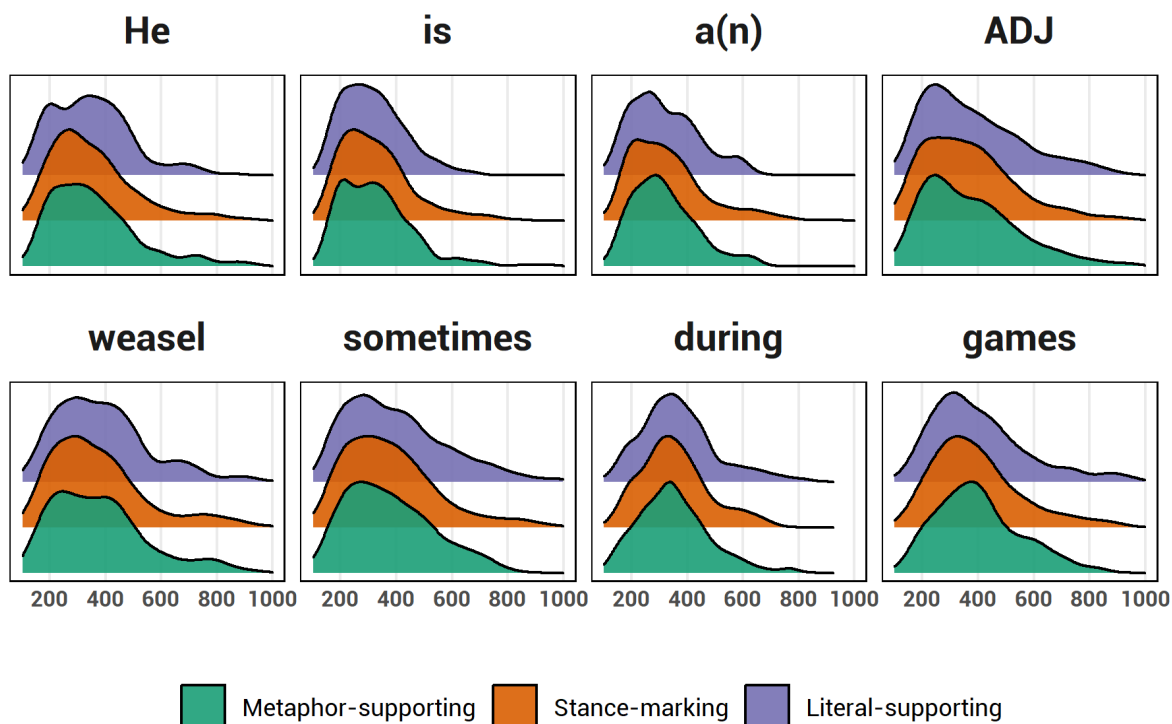
Figure 5.1 Reading time distributions (in milliseconds). Each panel shows the response distributions from an individual sentence region. *Metaphor-supporting* items are shown in green, *Stance-marking* in orange, and *Literal-supporting* in purple.

already be biased by the semantics of the adjective. While this picture looks qualitatively similar at both the noun and the critical adverb following it, the difference between the *Metaphor-supporting* and the *Literal-supporting* conditions is at its largest at the adverb *sometimes*, suggesting that comprehenders' online expectations vary as a function of incrementally combining semantic information from the adjective and the noun. Crucially, throughout these three regions, items in the *Stance-marking* condition are also read faster than items in the *Literal-supporting* condition. In what follows I describe these results in greater detail.

In order to quantify uncertainty over the results reported below, I fitted a Bayesian hierarchical model predicting log-transformed RTs at the critical region *sometimes* as a function of the experimental conditions, with the condition *Literal-supporting* dummy-coded

as the reference level. The model included the maximal random effect structure justified by the design, which in the present case is random slopes and intercepts for both participants and items. I model group-level variation within the items nesting them within the animal nouns, given that random variation at the level of the sentences should be accounted for both in terms of the critical mappings between the nouns and the adjectives as well as in terms of any idiosyncrasies tied to the individual animals used as vehicles in the metaphors. The model, fitted using the R package `brms` (Bürkner, 2017) and described in detail in the scripts available in the supporting material, had the following form, shown in `brms` syntax:

$$\log(\mathrm{RT}) \sim \text{condition} +$$
$$(1 + \text{condition} \mid \text{participant}) +$$
$$(1 + \text{condition} \mid \text{animal/ item})$$

For each highlighted result, I report the evidence provided by the statistical model in favor of the empirically attested differences (or lack thereof). In a Bayesian statistical framework, one is interested in the joint posterior distribution of the parameters of the model, which indicates a plausible range of values for the parameters given the model and the data at hand. I report a credibility interval (CrI) and the posterior probability that the parameter of interest, β, is smaller than zero ($P(\beta < 0)$). While this measure of evidence is gradient, one can speak of strong evidence for an effect when zero is not included in the CrI and $P(\beta < 0)$ is close to either zero or one, whereas when $P(\beta < 0)$ is around 0.5 one can speak of strong evidence for the absence of any effect. Concretely, I am interested in the difference between estimated values for cell means of the three experimental conditions, *Metaphor-supporting*, *Stance-marking*, and *Literal-supporting*.

The first comparison of interest is between the *Metaphor-supporting* and the *Literal-supporting* conditions. The account entertained here predicts that, in the posterior distribution of the Bayesian regression model, the difference in cell means $\beta_{\mathrm{H1}} = $ [estimates for cell mean of *Metaphor-supporting*] - [estimates for cell mean of *Literal-supporting*] should be

credibly smaller than zero, so that $P(\beta_{H1} < 0)$ should be large, i.e., very close to 1. This baseline prediction is based on the assumption that processing *sometimes* in a *Metaphor-supporting* scenario (*He is a devious weasel sometimes ...*) is less cognitively effortful than doing so in a *Literal-supporting* scenario (*He is a furry weasel sometimes ...*), as *sometimes* is more congruent with an unfolding interpretation which is more likely non-literal than literal, which in turn results in the adverb being more predictable in the *Metaphor-supporting* scenario. The second comparison of interest is between the *Stance-marking* and the *Literal-supporting* conditions. Crucially, there are two competing predictions here. If the stance-marking adjectives are interpreted as pragmatic cues to non-literality, the same pattern as above is expected, that is, the difference in cell means $\beta_{H2} =$ [estimates for cell mean of *Stance-marking*] - [estimates for cell mean of *Literal-supporting*] should be credibly smaller than zero, so that $P(\beta_{H2} < 0)$ should be large, i.e., very close to 1. This prediction is based on the hypothesis that processing *sometimes* in a *Stance-marking* scenario (*He is a real weasel sometimes ...*) is less cognitively effortful than doing so in a *Literal-supporting* scenario (*He is a furry weasel sometimes ...*), as the stance-marking adjectives are interpreted as intensificational cues which support the likely non-literal interpretation cued by the noun. If, on the other hand, the adjectives are interpreted at face value in terms of their truth-conditional meaning, i.e., as indicating that a true or real weasel is being talked about, then the difference in cell means $\beta_{H2} =$ [estimates for cell mean of *Stance-marking*] - [estimates for cell mean of *Literal-supporting*] should *not* be credibly smaller than zero, so that $P(\beta_{H2} < 0)$ should not be large, i.e., not close to 1.

Visual inspection of the results in Figure 5.2 suggests that, at the *sometimes* region, items in the *Metaphor-supporting* condition are read faster than items in the *Literal-supporting* condition, in line with the baseline prediction. This result finds relatively strong quantitative support in the respective statistical model ($\beta_{H1} =$ -0.04, 95% CrI [-0.09, 0.00], $P(\beta_{H1} < 0) =$ 0.94). As for the items in the *Stance-marking* condition, the descriptive results suggest that, much like items in the *Metaphor-supporting* condition, they are read faster than items in the
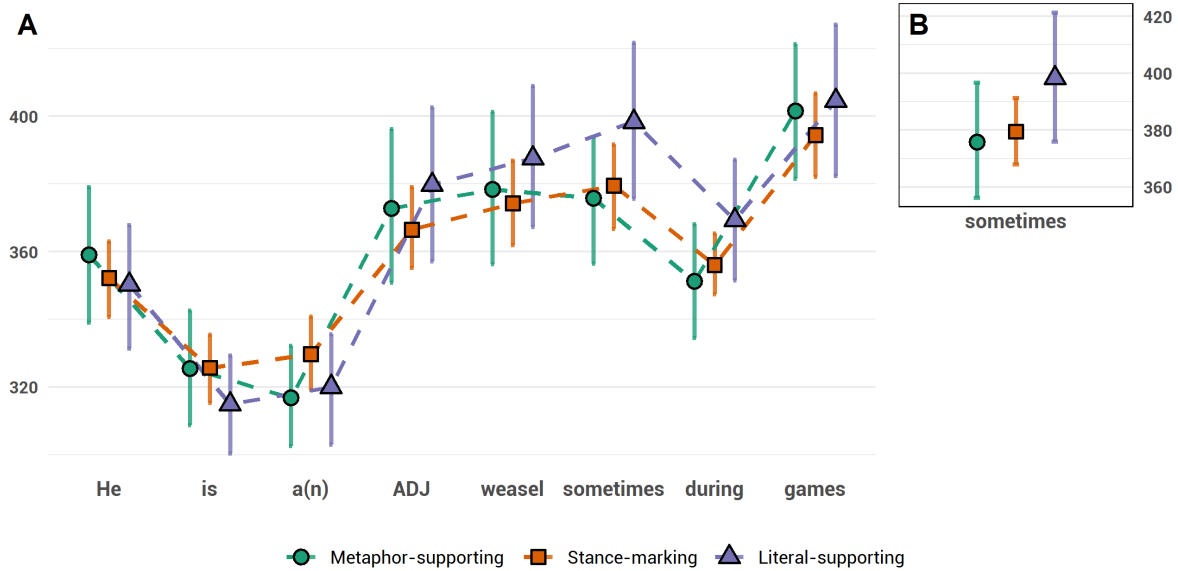
Figure 5.2 Panel A shows the mean reading times at each sentence region. Panel B shows a close-up of the mean reading times at the adverb region. *Metaphor-supporting* items are shown in green, *Stance-marking* in orange, and *Literal-supporting* in purple. The error bars represent 95% bootstrapped confidence intervals.

Literal-supporting condition, in line with the prediction derived from the hypothesis of stance markers as predictive pragmatic cues. This result finds particularly strong quantitative support in the respective statistical model (β_{H2} = -0.04, 95% CrI [-0.08, 0.00], $P(\beta_{\text{H2}} < 0) = 0.96$). The descriptive results further suggest no reliable difference between the items in the *Stance-marking* and *Metaphor-supporting* conditions (see regression coefficients in Table 5.1 below).

Region	Hypothesis	β	95%-CrI	P($\beta < 0$)
adverb	Metaphor-supporting < Literal-supporting	-0.04	[-0.09, 0.00]	0.94
adverb	Stance-marking < Literal-supporting	-0.04	[-0.08, 0.00]	0.96
adverb	Metaphor-supporting < Stance-marking	0.00	[-0.05, 0.04]	0.55

Table 5.1 Regression coefficients.

Exploratory analyses

Stance markers

Recall that each item in the *Stance-marking* condition contained one of three different adjectives, namely *actual*, *literal*, or *real*. In order to explore whether there are systematic processing differences depending on the particular intensifier participants read, I present the same data as in Figure 5.2 plotting *actual*, *literal*, and *real* separately. In order to quantify uncertainty over the empirical results, I again fitted a hierarchical model predicting log-transformed RTs as a function of the different experimental conditions, splitting the *Stance-marking* condition into three separate groups, each corresponding to one of the stance-marking tokens, and dummy-coding the *Literal-supporting* condition as the reference level. Figure 5.3 below shows the raw RTs according to the five modifier groups.

Visual inspection of the results suggests that there is potentially systematic variation across the three tokens in the *Stance-marking* condition. Crucially, while the descriptive results suggest that all three tokens are read faster than items in the *Literal-supporting* condition, taken as separate groups in themselves, only in the case of *literal* items there is relatively strong evidence for a reliable mean reading time difference compared to *Literal-supporting* items (see regression coefficients in Table 5.2 below). The descriptive results further suggest no reliable difference between the items in any of the three *Stance-marking* groups and those in the *Metaphor-supporting* condition.
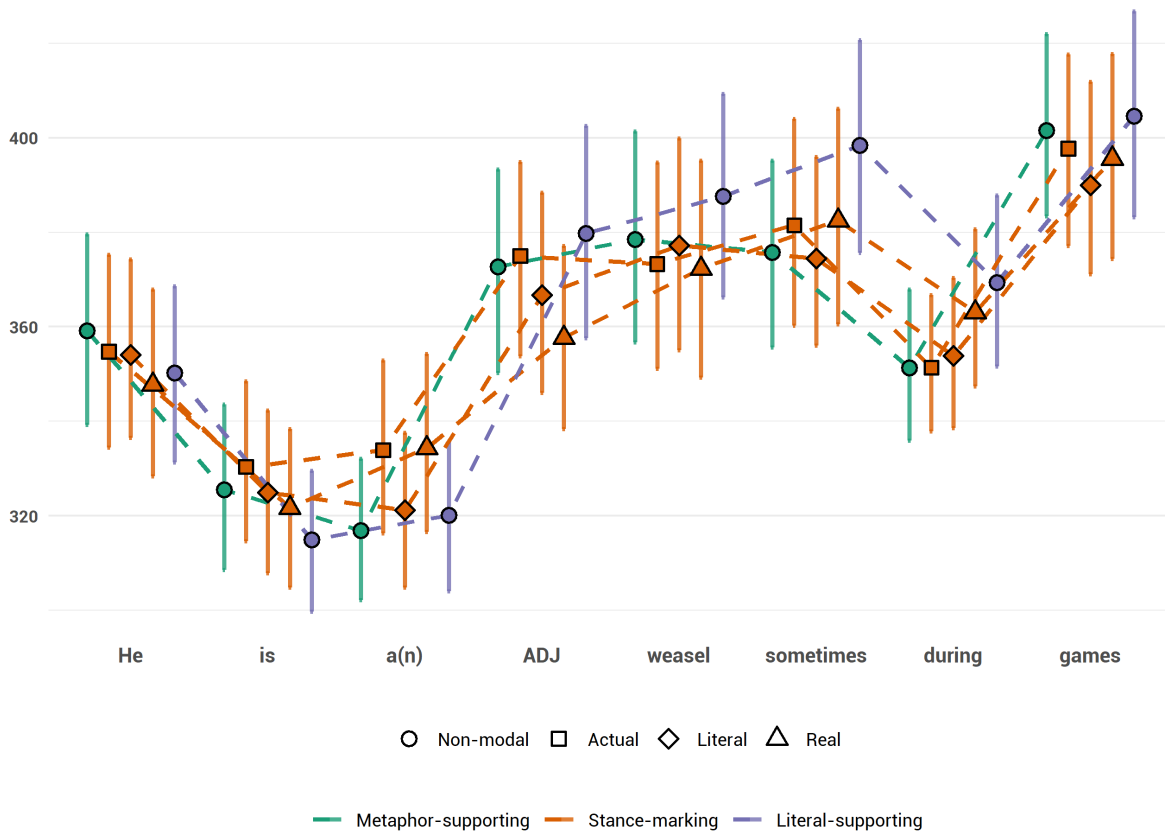
Figure 5.3 Mean reading times at each sentence region. *Metaphor-supporting* items are shown in green, *Stance-marking* in orange, and *Literal-supporting* in purple. Squares indicate *Actual* items, diamonds indicate *Literal* items, and triangles indicate *Real* items. The error bars represent 95% bootstrapped confidence intervals.

Region	Hypothesis	β	95%-CrI	P($\beta < 0$)
adverb	Metaphor-supporting < Literal-supporting	-0.04	[-0.09, 0.01]	0.93
adverb	Actual < Literal-supporting	-0.03	[-0.08, 0.01]	0.87
adverb	Literal < Literal-supporting	-0.05	[-0.11, 0.00]	0.94
adverb	Real < Literal-supporting	-0.04	[-0.09, 0.01]	0.89
adverb	Metaphor-supporting < Actual	-0.01	[-0.07, 0.04]	0.66
adverb	Metaphor-supporting < Literal	0.01	[-0.05, 0.07]	0.42
adverb	Metaphor-supporting < Real	0.00	[-0.06, 0.05]	0.55
adverb	Actual < Literal	0.02	[-0.04, 0.08]	0.27
adverb	Actual < Real	0.01	[-0.05, 0.06]	0.40
adverb	Literal < Real	-0.01	[-0.07, 0.05]	0.64

Table 5.2 Regression coefficients.

5.3 Discussion

In the present study, I set out to test whether adjectives with a conventionalized pragmatic meaning of intensification act as predictive cues to non-literality in the online processing of language. I investigated this question in English using a word-by-word self-paced reading experiment where participants read metaphoric statements of the sort *He is a real weasel sometimes during games*, where the adjective *real* acts as a marker of stance, more specifically as an intensifier of the metaphoric phrase *He is a weasel*. I compared the reading times measured at the adverb *sometimes*, which by design directly followed the metaphoric phrase and was meant to constrain the unfolding interpretation of the sentence in terms of a likely non-literal meaning. Crucially, I compared the reading times across three experimental conditions: the *Stance-marking* condition (*He is a real weasel sometimes during games*), the *Metaphor-supporting* condition (*He is a devious weasel sometimes during games*), and the *Literal-supporting* condition (*He is a furry weasel sometimes during games*). While the *Stance-marking* condition contained the critical intensifying adjectives, the *Metaphor-supporting* and *Literal-supporting* conditions contained non-intensifying adjectives which, given their lexical semantics, were more likely to support either a metaphoric or literal interpretation of *He is a weasel*, respectively.

The hypothesis I entertained was that the intensifying adjectives would be interpreted as pragmatic cues to non-literal meaning, such that they would reinforce the meaning of a metaphor like *He is a weasel* and thus behave qualitatively similar to adjectives with a semantics which cues a more metaphoric interpretation and, by extension, unlike adjectives with a semantics which cues a more literal interpretation. In more concrete terms, participants were expected to read the critical adverb *sometimes* faster in the *Stance-marking* condition than in the *Literal-supporting* condition, no systematic reading time differences being expected between the *Stance-marking* and the *Metaphor-supporting* conditions. Participants were also expected to read the critical adverb *sometimes* faster in the *Metaphor-supporting*

condition than in the *Literal-supporting* condition.

The results show that, as predicted by an account under which intensifying adjectives serve as pragmatic cues to non-literality despite their truth-conditional lexical meaning, participants read items in the *Stance-marking* condition qualitatively similar to items in the *Metaphor-supporting* condition, such that there are no reliable mean reading time differences between the two conditions at the critical adverb region. The results also show that, as predicted by the abovementioned account, participants read items in the *Stance-marking* condition faster than items in the *Literal-supporting* condition, such that there are reliable mean reading time differences between the two conditions at the adverb. Finally, as predicted more generally by an expectation-based account of processing, participants read items in the *Metaphor-supporting* condition faster than items in the *Literal-supporting* condition, such that there are reliable mean reading time differences between the two conditions at the adverb, despite evidence for such an effect being slightly weaker compared to the comparison between the *Stance-marking* and *Literal-supporting* conditions.

In addition to these main results, which follow from a confirmatory analysis of the data, I reported an exploratory analysis aimed at investigating any potential systematic differences between the three tokens in the *Stance-marking* condition, namely *actual*, *literal*, and *real*. As these results show, only in the case of *literal* items there is relatively strong evidence for a reliable mean reading time difference compared to the items in the *Literal-supporting* condition, such that the former are read, on average, faster than the latter. However, it is worth noting that, despite finding weaker evidence in the statistical model, the descriptive results suggest that *actual* and *real* items behave, on average, more like *Metaphor-supporting* items than *Literal-supporting* ones, which is to be expected given the results based on the aggregated data. Since the exploratory analysis involved partitioning the items in the *Stance-marking* condition and modeling them as separate groups, power was severely reduced as each token only occurs once or twice per participant compared to five occurrences for items in the *Literal-supporting* and *Metaphor-supporting* conditions. As such, it might be fair to

assume that there are systematic differences between the tokens, although such differences cannot be reliably quantified using the present data.

Taken together, the results suggest that intensifying adjectives can act as pragmatic cues in the incremental interpretation of language, at least in the processing of English. In particular, the results highlight how such markers of stance serve as *predictive* cues to non-literal meaning, such that in the processing of non-literal, metaphoric statements they cue a likely non-literal interpretation despite their lexical semantics related to truth or reality. This is in line with the hypothesis that, given their function as markers of subjective stance, *actual*, *literal*, and *real* were more likely to be interpreted in terms of their intensificational meaning than in terms of their lexical semantic meaning. Underlying this hypothesis was the assumption that, as a cognitive process, interpretation involves dynamically weighing the cues available in a communicative signal, such that at any given moment in processing the most likely outcome of the interpretive process consists of the combined weight of all cues processed thus far (Rohde and Kurumada, 2018). An auxiliary assumption was that interpretation is incremental in nature: incoming cues are weighted as they are processed, their functional weight depending, among other things, on their conditional likelihood in a particular linguistic and discursive context (Degen and Tanenhaus, 2019; Rohde and Kurumada, 2018).

In terms of implications for accounts of (incremental) interpretation and (predictive) processing, the current results add to the body of evidence suggesting that the relative weight of an adjective – or any other linguistic item, for that matter – varies depending on its functional relevance for interpretation in context (Gardner et al., 2021; Rubio-Fernandez and Jara-Ettinger, 2020; Tourtouri et al., 2021). Crucially in the case of adjectives with a conventionalized meaning of intensification, such functional relevance seems to depend on some weighing between both their lexically encoded semantic meaning and their conventionalized pragmatic meaning. In fact, as discussed in Section 5.1.1, different modifiers impart more or less of their lexical meaning to their conventionalized meaning as stance markers, such that

in the case of English adjectives one cannot straightforwardly distinguish one meaning from the other on the basis of cues from linear ordering alone. This raises a number of questions about the putative causal mechanisms underlying such weighing.

Causally, the process of weighing the meaning of a polysemous/ polyfunctional lexical item like *real* can be assigned both proximal and distal explanations. From a more distal, evolutionary perspective, the availability of a given conventionalized pragmatic meaning can be said to be tied to the synchronic stage at which the lexical item finds itself on the relevant pragmaticalization process (Diewald, 2011; Traugott, 2012, 1995), such that its weight in any particular usage event varies as a function of how far along the item is on a diachronic pathway to becoming a (pragmatic) stance marker. This might explain the marker's meaning distribution at any given moment in sychronic time, including any pragmatic meanings it might have, however, it does not explain how the relevant meaning is actually activated in interpretation. Thus, from a more proximal, psycholinguistic perspective, it is reasonable to assume that certain conventionalized markers have more of a weight as predictive cues than others, likely due to a combination of both functional and formal properties such as their relative frequency in discourse and their relation to other items in the lexicon/ grammar, including items with which they might share similar functions and/ or forms. From such a psycholinguistic perspective, which is the level of analysis this study is concerned with, one must assume, mechanistically, that some (cognitive) process allows comprehenders to activate the most likely meaning representation in their mental lexicons. In the case at hand, a number of heuristics could be said to map onto such a process.

At first sight, reasoning about the (Gricean) quality of the utterance might explain the interpretive process: saying that someone is a *real weasel* is blatantly false, and so by virtue of the proposition being explicitly not true it likely means that the utterance is meant non-literally. However, while this might explain the interpretation of a construction where the modifier straightforwardly supports the metaphoric meaning of the noun given its semantic meaning, like *devious weasel*, it does not explain how the intensifying meaning of *real* takes

precedence over its lexical meaning of 'reality', at least not without additional assumptions. An alternative account would be one grounded in reasoning about conventional implicatures (Potts, 2012, 2003), such that the conventionalized meaning of *real* as an intensifier supports the metaphoric interpretation of *real weasel*. This seems to provide the most parsimonious explanation for the outweighing of the pragmatic meaning of *real* with relation to its lexical semantic meaning: interpretation involves some process of selection between the two conventionalized meanings, even if the intensifying meaning itself might be conventionalized to a lesser extent than its truth-conditional counterpart. An alternative heuristic which might be said to account for the interpretation at hand is one which assigns *real* a role as a procedural cue. In line with this view, certain, if not all, linguistic expressions provide cues to interpretation insofar as they encode instructions as to how to process the relevant proposition(s) in an utterance (Blakemore, 2002, 1987), either independently from or in addition to contributing to the truth-conditions of those same propositions (Carston, 2016). In the current case, one can thus assume a procedural role for *real*, such that it cues how to interpret the unfolding proposition *He is weasel sometimes during games*. However, both the conventional implicature account and the procedural account do not specify what exact process allows the computation of *real* primarily as a (pragmatic) procedure as opposed to a conceptual representation, which means that the two account are, all else equal, equally parsimonious. Ultimately, then, in order to more satisfactorily account for the actual process underlying meaning selection in the interpretation of *real*, only a more detailed algorithmic specification of the putative computational principles might help clarify whether one account is more parsimonious than the other.

A further issue concerning the interpretation of conventionalized stance markers like *real* has to do with how interpretation is jointly constrained by several synchronic sources of variation, from (lexical) semantics and pragmatics to syntax and prosody, to consider canonical linguistic sources only. An outstanding theoretical question at this junction relates to how such manifold sources of variation are integrated in processing, and how exactly cue

weighing operates when information is derived from multiple sources, which is the norm in naturalistic processing in the wild. As discussed before, the adjectival modifiers investigated in the present study occur in restricted syntactic configurations, namely before the modified element, whether that is a noun (e.g. *He is a real weasel*) or another adjective (e.g. *He is real tall*). Their adverbial counterparts in English, on the other hand, can have different meanings depending, among other things, on their syntactic position. For instance, while in *He really is a weasel* the adverb signals a subjective emphasis on the truth of the proposition at hand, i.e., that the referred person is a metaphoric weasel, in *He is really a weasel* the same adverb seems to reinforce the weasel-like character of the referred person. In light of such subtle meaning differences, it is unclear, all else being equal, in which context *really* might have a stronger weight as a predictive cue to non-literality, and whether any difference in strength varies as a function not only of *really*'s meaning as a stance-marking cue but also of its syntactic position. Future research could investigate how such fine-grained variations in meaning, resulting at the very least of an interaction between semantics/ pragmatics and syntax, can impact (incremental) interpretation and cue weighing in online processing.

In terms of implications for accounts of semantics and pragmatics more generally, although the results of the exploratory analysis do not show robust systematic differences between the different stance-marking tokens in the present study, one might speculate about any potential differences between *actual*, *literal*, and *real*. Indeed, as discussed earlier, it might be that the three intensifying adjectives find themselves along different stages of their pragmaticalization process. While all three modifiers share a similar lexical semantic meaning related to truth or reality, as intensifiers, some might be more conventionalized than others, which might impact their functional weight when being interpreted in online processing. A similar situation might be expected for their adverbial counterparts in English, where *actually* has a much broader and more polyfunctional meaning profile than *really* and *literally* (Maschler and Schiffrin, 2015). As discussed earlier, when comparing adjectives and adverbs, such differences in their range of meanings might be due to formal as well as

distributional properties, whereby, at least in English, adverbs are syntactically more flexible than adjectives and thus might be more susceptible to developing more varied and nuanced meaning profiles. Exploring the interactions between different lexical categories and their potential in expressing various meanings is an exciting avenue for future (psycho)linguistic research, especially research which seeks to establish explicit links between diachronic meaning development and interpretation.

Methodologically speaking, all of the conclusions and speculations discussed so far are based on the critical comparisons drawn between the processing signatures attested in the different experimental conditions, which included items designed to be either more or less congruent with a non-literal interpretation of a given metaphoric sentence. While the experimental items were designed to differ from one another in systematic and categorical ways, the particular mappings between the individual adjectives and nouns in each condition were idiosyncratic in nature, such that certain combinations of a metaphor vehicle with a particular adjective are more common than others. Although these combinations were designed with collocational concerns in mind, and were modeled in terms of any idiosyncratic behavior, natural variation within the experimental conditions was expected. This within-condition variation could be observed in the variability around the condition means seen in the data in Fig. 5.2. While there was considerable variability within the non-intensifying conditions, this does not undermine the validity of the inferences drawn from the data in Section 5.2. Nevertheless, it is a limitation of the present setup which deserves acknowledgment, and which could be more carefully controlled in similar setups, for instance, by establishing the relative frequencies of the critical adjective-noun collocations, which by extension would result in more balanced surprisal measures for the critical adjective-noun pairs. However, as it is, the design and the setup of the current study provide a solid first step in investigating an underexplored phenomenon which, by definition, is hard to control experimentally while maintaining the naturalness of the stimuli.

All in all, the present study sheds light on the potential mechanisms underlying the

interpretation of lexical pragmatic markers such as the intensifying adjectives *actual*, *literal*, and *real* in English. On the face of it, interpreting an utterance like *He is a real weasel* involves weighing both the straightforward literal meaning of *real* as a lexical item and its conventionalized pragmatic meaning as a marker of (epistemic) stance. Deriving therefrom the most likely interpretation in context can be said to involve either drawing an inference on the basis of the adjective's conventionalized (pragmatic) meaning or processing the word as a procedural cue to meaning. Without further specifying what the exact processes underlying interpretation might be, the two accounts, as such, seem to hold comparable explanatory power in capturing the mechanism at play behind the interpretation of a marker like *real*. More generally, a fruitful avenue for future research might be to consider how different dynamics and sources of variation jointly constrain the interpretive mechanisms involved in the incremental activation of lexical meaning and the pragmatic processing of language.

REFERENCES

Aikhenvald, Alexandra Y (2004). *Evidentiality.* OUP Oxford.

— (2012). "The essence of mirativity". In: *Linguistic Typology* 16.3, pp. 435–485.

Athanasiadou, Angeliki (2007). "On the subjectivity of intensifiers". In: *Language sciences* 29.4, pp. 554–565.

Bardenstein, Ruti and Mira Ariel (2021). "The missing link between truth and intensification". In: *Studies in Language.*

Blakemore, Diane (2002). *Relevance and linguistic meaning: The semantics and pragmatics of discourse markers.* Vol. 99. Cambridge university press.

— (1987). "Semantic constraints on relevance". In:

Boye, Kasper (2012). *Epistemic meaning.* De Gruyter Mouton.

Bürkner, Paul-Christian (2017). "brms: An R package for Bayesian multilevel models using Stan". In: *Journal of statistical software* 80.1, pp. 1–28.

Carston, Robyn (2016). "The heterogeneity of procedural meaning". In: *Lingua* 175, pp. 154–166.

Cornillie, Bert (2009). "Evidentiality and epistemic modality: On the close relationship between two different categories". In: *Functions of language* 16.1, pp. 44–62.

Cornillie, Bert and Paola Pietrandrea (2012). "Modality at work. Cognitive, interactional and textual functions of modal markers". In:

de Haan, Ferdinand (2012). "Evidentiality and mirativity". In: *The Oxford handbook of tense and aspect*, pp. 279–308.

Haan, Ferdinand de (2001). "The relation between modality and evidentiality". In: *Linguistische Berichte* 9, pp. 201–16.

Degen, Judith and Michael K Tanenhaus (2019). "Constraint-based pragmatic processing". In: *The Oxford handbook of experimental semantics and pragmatics*, pp. 21–38.

Diewald, Gabriele (2011). "Pragmaticalization (defined) as grammaticalization of discourse functions". In: *Linguistics* 49.2, pp. 365–390. DOI: doi:10.1515/ling.2011.011.

Gardner, Bethany et al. (2021). "Online pragmatic interpretations of scalar adjectives are affected by perceived speaker reliability". In: *PloS one* 16.2, e0245130.

Ghesquierè, Lobke (2017). "Intensification and focusing". In: *Exploring intensification: Synchronic, diachronic and cross-linguistic perspectives* 189, p. 33.

Grandi, Nicola and Lívia Körtvélyessy. *Edinburgh handbook of evaluative morphology.*

Gray, Bethany and Douglas Biber (2012). "Current conceptions of stance". In: *Stance and voice in written academic genres.* Springer, pp. 15–33.

Grodner, Daniel and Julie C Sedivy (2011). "10 The Effect of Speaker-Specific Information on Pragmatic Inferences". In: *The processing and acquisition of reference.* Vol. 2327. MIT Press, pp. 239–272.

Heine, Bernd, Tania Kuteva, Heine Bernd, et al. (2002). *World lexicon of grammaticalization.* Cambridge University Press.

Jegerski, Jill and Bill vanPatten (2013). *Research methods in second language psycholinguistics.* Routledge.

Körtvélyessy, Lívia (2015). *Evaluative morphology from a cross-linguistic perspective.* Cambridge scholars publishing.

Maschler, Yael and Deborah Schiffrin (2015). "Discourse markers: Language, meaning, and context". In: *The handbook of discourse analysis* 1, pp. 189–221.

Naylon-Roberts, Ffion (2016). "Processing Modified Metaphors: A real pain in the neck?" MA thesis.

Paradis, Carita (2012). "Between epistemic modality and degree: the case of really". In: *Modality in contemporary English.* De Gruyter Mouton, pp. 191–222.

Potts, Christopher (2012). "94. Conventional implicature and expressive content". In: *Volume 3.* De Gruyter Mouton, pp. 2516–2536.

— (2003). "Expressive content as conventional implicature". In: *Proceedings-Nels.* Vol. 33, pp. 303–322.

Ricci, Claudia and Corinne Rossari (2018). *Commitment phenomena through the study of evidential markers in Romance languages.*

Rohde, Hannah and Chigusa Kurumada (2018). "Alternatives and inferences in the communication of meaning". In: *Psychology of Learning and Motivation.* Vol. 68. Elsevier, pp. 215–261.

Rubio-Fernandez, Paula and Julian Jara-Ettinger (2020). "Incrementality and efficiency shape pragmatics across languages". In: *Proceedings of the National Academy of Sciences* 117.24, pp. 13399–13404.

Sedivy, Julie C et al. (1999). "Achieving incremental semantic interpretation through contextual representation". In: *Cognition* 71.2, pp. 109–147.

Tourtouri, Elli N, Francesca Delogu, and Matthew W Crocker (2021). "Rational Redundancy in Referring Expressions: Evidence from Event-related Potentials". In: *Cognitive Science* 45.12, e13071.

Traugott, Elizabeth Closs (2012). "Pragmatics and language change". In: *The Cambridge handbook of pragmatics*, pp. 549–566.

— (1995). "Subjectification in grammaticalization". In: *Subjectivity and subjectivisation: Linguistic perspectives* 1, pp. 31–54.

Wu, Shiyu and Zheng Ma (2020). "How is Chinese reading affected by under-specification and over-specification? Evidence from self-paced reading experiments". In: *Journal of Pragmatics* 155, pp. 213–233.

Chapter Six

Pragmatic inferencing in incremental processing as a window into the mental(istic) processes of language use

6.1 Of pragmatic cues to meaning and expectations

In this dissertation, I've presented four empirical studies focused on phenomena related to incremental pragmatic interpretation. Chapter 2 introduced the issue of pragmatic inferencing in incremental processing in the context of a well-studied pragmatic phenomenon, namely the interpretation of utterances with scalar expressions, more specifically scalar quantifiers in German. Chapter 3 shifted the focus to the main theme of the dissertation, namely the incremental interpretation of utterances containing markers of epistemicity, focusing more specifically on the interpretation of negated polar questions in English and German with morphosyntactic forms which signal different epistemic stances. Chapter 4 extended the focus on epistemicity markers to the interpretation of utterances with lexical markers of stance, more specifically stance adjectives in English which have a conventionalized pragmatic meaning of intensification alongside their lexical semantic meaning. Finally, Chapter 5 concluded the empirical coverage of the thesis, focusing on the interpretation of utterances

with particles in German which signal common ground relations of contrast and agreement. In what follows, I lay the findings side by side and relate them to one another, focusing on the three main issues this dissertation is concerned with, as outlined in Chapter 1: how comprehenders might inference pragmatically via different types of implicatures, what role prediction and predictability might play in online pragmatic processing, and how indexical pragmatic meanings related to the marking of epistemicity might modulate interpretation incrementally and thus affect any actual inferences drawn during processing.

6.2   Inferencing pragmatically via implicatures

Across the four studies reported in chapters 2 to 5, I have investigated three types of pragmatic inferences: scalar implicatures, manner implicatures, and conventional implicatures. Common to all these types of inference is the underlying assumption that language users reason about the usage of linguistic forms in context, whether in terms of the informativity of a form vis-à-vis its scalar counterparts, the felicity of a given utterance formulation in light of epistemic assumptions, or the felicity of a given pragmatic meaning conventionally associated with a lexical form. Beyond this common core, certain auxiliary assumptions are common, or at least can be applied similarly, to particular pairings of these inferences: the assumption that language users reason about the usage of alternatives, whether alternative lexical forms or alternative utterance formulations, is common to both scalar and manner inferences, while the assumption that a meaning conventionally associated with a linguistic form can give rise to a pragmatic inference is common to both manner and conventional inferences. Moreover, depending on how the particular phenomenon under consideration is characterized, the assumption that the link between a meaning and a form is indexical might also apply to both manner and conventional inferences, if not to pragmatic interpretation more generally.

Alternative-based inferencing

When it comes to the role of alternatives in pragmatic inferencing, insofar as the phenomena investigated in this thesis are concerned, it is clear that scalars and a theory of scalarity can be said to affect inferencing in qualitatively similar ways as non-scalar alternatives. The degree to which they do so, however, might be different. Lexical scales such as one containing the quantifiers *some* and *all* provide a set of linguistic alternatives in their own right (Horn, 1985), one which might be pragmatically relevant in context if the relationship between two or more scalemates makes a contextual element more salient than another (Gotzner and Romoli, 2022). For instance, the usage of the quantifier *some* in a referring expression like *Some of the triangles are blue* might make a referential contrast between two potential referents salient given that the interpreter might reason that one of the referents is best referred to using the semantically stronger alternative *all*, making the lexical choice pragmatically relevant for interpretation such that there might be enough of a processing pressure for the interpreter to draw a scalar inference.

Alternative utterance formulations, on the other hand, do not constitute a set of of linguistic alternatives in themselves, although this might vary depending on the nature of the alternatives and the meanings associated with them. As discussed in Chapter 3, the features which vary between two alternative utterance formulations can be of any linguistic nature: for instance, both prosodic and syntactic modulations of a sentence can serve as cues to intended meaning in context, whereby both a pitch accent and a particular word order can give rise to a contrastive inference (e.g. Kaiser and Trueswell, 2004; Kurumada et al., 2014). While linguistic scales are not restricted to lexical scales, the nature of the relationship between any two items on such a scale is primarily semantic – regardless of whether the scale denotes quantities, degrees of goodness as on a scale containing the adjectives *good* and *bad*, or any other semantic dimension. Thus, by virtue of their semantic relationship, alternatives on a linguistic scale seem more likely to be pragmatically relevant for interpretation than

193

alternatives which are not ordered on any scale according to their meaning. Indeed, unlike scalar alternatives, the relationship between non-scalar alternatives might be primarily formal or distributional in nature, the meanings that are associated with such alternatives not standing in any natural order relative to one another. For instance, whether a verb receives a contrastive focus accent might be, from an interpretive perspective, not as strong a pragmatic cue as using a scalar term like *some*, meaning that, abstracting away from other factors, the latter might be more likely to make a relevant discourse element salient and thus to give rise to an inference on grounds of semantic salience alone. Of course, comparing one inference type to another on the basis of linguistic cues of different sorts is an unfit comparison, but it nevertheless highlights how a graded strength of interpretation seems to underlie the derivation of different types of pragmatic inferences. While any linguistic feature in their own right – prosodic, lexical or otherwise – might provide strong enough of a pragmatic cue for interpretation in context, the nature of the inference to be drawn might affect the likelihood of activating the inferential process in the first place (a point which is further discussed in section 6.3 below). The linguistic forms discussed in Chapter 3 are an interesting case study in this regard.

Unlike scalars, alternative morphosyntactic forms of a question like *Didn't you buy the ticket?* and *Did you not buy the ticket?* are not conceived as being ordered on any scale according to their semantics. Yet, unlike other non-scalar alternatives which are known to give rise to inferences, like a contrastive vs. a non-contrastive pitch accent, each of these forms carries semantic meaning in itself, which is to say that their relationship to one another can be conceived of as being semantic in nature, similarly to the relationship between, for instance, the quantifiers *some* and *all*. Perhaps more interestingly, the nature of the relationship between these different alternatives can also be conceived of as being pragmatic: the pairs <*Didn't you buy the ticket?*, *Did you not buy the ticket?*> and <*Hast du kein Ticket gekauft?*, *Hast du nicht ein Ticket gekauft?*> can be ordered according to the relative sense of epistemic confidence they express, where what varies between one and the

other form is the degree to which the utterer commits to the truth of the content expressed in the utterance. How does this possibly impact inferencing compared to inferencing via scalars?

As discussed in Chapter 3, in the case of questions with biases, pragmatic felicity – and by extension the likelihood of drawing a pragmatic inference – is construed as being dependent on assumptions about the epistemic state of the utterer. In other words, unlike in the case of scalars, where the core inferential assumption is one that the utterer is cooperative and therefore uses scalar expressions informatively, in the case of questions with biases there is a more fine-grained set of assumptions about what exact linguistic forms map more felicitously onto information which the interpreter might be uncertain about. It might be reasonable to conclude then that, from a processing perspective, inferencing on the basis of alternative utterance formulations as in alternative morphosyntactic forms of a negated polar question might require stronger evidence to activate the inferential process compared to inferencing on the basis of lexical scalars[1], despite both seemingly involving a process of abduction where the relevant evidence is derived from a combination between a linguistic cue and extra-sentential contextual information. Indeed, as discussed in Chapter 3, the conclusion that non-scalars alternatives generally act as weak inferential triggers is motivated by the finding that only in principle does the felicity of the mapping between a question form and its epistemic meaning actually impact incremental interpretation, which implies even a lower likelihood of such question forms serving as pragmatic cues to meaning as is commonly assumed in the literature. While the type of discourse move these particular forms are used in must play a non-trivial role in the inferencing process – questions pose different processing demands compared to statements (see e.g. Levinson, 2016; Stivers et al., 2009) – it is nonetheless theoretically relevant to ask whether manner implicatures, generally speaking, require more evidence to be triggered compared to scalar implicatures (see Gotzner, Solt, et al. (2018) for

---

[1]Though, as discussed in the next section, the *type* of meaning a given form signals is likely to be a more relevant factor in the derivation of a pragmatic inference than the actual nature of any alternative.

a comparison between scalar and manner inferences triggered by scalar alternatives).

Ultimately, based on the phenomena tackled in chapters 2 and 3, there seem to be both theoretical and empirical reasons to support a view of conversational implicatures, and perhaps of pragmatic inferencing more generally, as a process based on graded strength of interpretation, a position which has been explicitly argued for, e.g., by Ariel (2016)[2] and Macagno (2012). At the same time, given the results reported in Chapter 2 where predictable linguistic information was found to be harder to process after processing scalars which were informative in context, it seems reasonable to conclude that such a gradience in the likelihood of drawing an inference is linked not only to the core theoretical assumptions commonly associated with each particular inference type but perhaps more importantly to a larger set of constraints at play during interpretation, one which seems to involve processing demands which act as stronger constraints as those expected to be involved in the interpretive process as per the assumptions traditionally discussed in the literature.

Indeed, as seen in Chapter 2, one such demand can be a pressure to integrate crucial semantic information incrementally during processing, whereas, as seen in Chapter 3, a related demand can be one to generate, as quickly as possible, a minimally coherent interpretation model, a pressure which can outweigh the derivation of secondary, non-truth-conditional meaning from forms which also encode truth-conditional meaning in the same utterance context. This conclusion is further supported by findings from work on the processing of adjective-modified referring expressions which have shown that what might be naively understood as pragmatic infelicity, namely overinformativeness, can actually be parsimoniously treated as pragmatic felicity when recasting felicity in terms of efficiency-based concerns as opposed to informativity-based ones (Aparicio et al., 2018; see also Tourtouri et al., 2021;

[2]In Ariel's terms, both the scalar and manner implicatures discussed here might be regarded as cases of communicated interpretations, where the pragmatically relevant meaning is said to be derived from the explicit content of the utterance. It is unclear, however, whether in her typology of pragmatic interpretations one would be predicted to be more likely to be interpreted pragmatically than the other.

Wu and Ma, 2020). Particularly relevant for this line of argumentation is the finding that the type of adjective used in a referring expression is crucial as to whether it serves as predictive pragmatic cue or not, which not only suggests a varied degree of pragmatic import when it comes to otherwise structurally similar linguistic forms but also points to the likely underlying explanatory dimension, namely whether a pragmatically relevant meaning is, in itself, more or less context-dependent. As discussed in the introduction, under the assumption that meanings can be construed as lying on a continuum from more truth-conditional to less truth-conditional, less truth-conditional meanings should be, by definition, more context-dependent and thus potentially less constraining than their more truth-conditional counterparts. One may find evidence of such gradience in pragmatic import both when comparing forms from the same formal class, e.g., when comparing color (e.g. *blue*) or size adjectives (e.g. *tall*) to quality adjectives (e.g. *unbroken*), as well as when comparing forms from formally distinct classes, e.g., when comparing stance adjectives to modal particles. In either case, the general prediction is that the less truth-conditional the meaning of a given linguistic form – either in comparison to other structurally similar forms or to altogether different linguistic forms with similar functions – the less likely it is to give rise to a pragmatic interpretation.

This relationship can be captured in a theoretical space as illustrated in Figure 6.1. As suggested by the figure, the relationship between pragmatic import and non-truth-conditionality is not expected to be linear, such that, on part of the space, forms that contribute meanings which are less truth-conditional are expected to have a stronger relative impact on interpretation compared to forms that contribute meanings which are more truth-conditional, whereas, on another part of the space, the reverse is true. As discussed in the next section, this relationship is expected not to be linear given that it is modulated by an additional factor, namely whether the relationship between a meaning and its associated form is indexical.
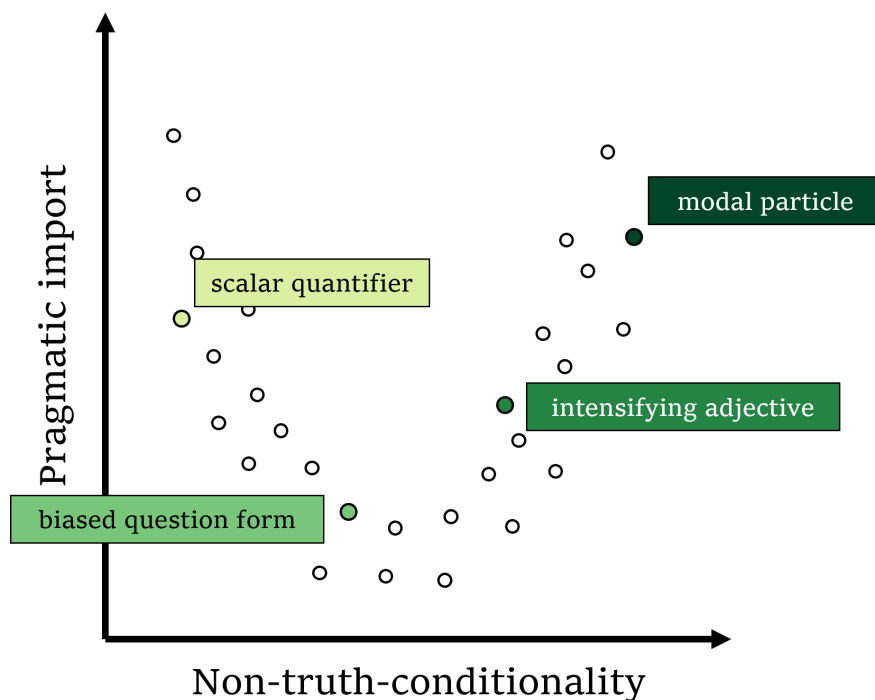
Figure 6.1 Hypothesized relationship between the pragmatic import of a given linguistic form and the nature of its meaning contribution in context. Each point represents one potential form/ phenomenon in such a theoretical space. The colored points represent the approximated location of the phenomena investigated in this dissertation.

Inferencing on conventional (pragmatic) meanings

When it comes to the role of conventional meaning in pragmatic inferencing, insofar as the phenomena investigated in this thesis are concerned, one general conclusion is that inferencing is modulated not only by the nature of the meanings conventionally associated with a particular linguistic form but also by their degree of conventionalization. As discussed in the introduction, gradience in meaning conventionalization is taken to mean variability in the extent to which a linguistic form is conventionally associated with a given meaning. So, if a form-meaning mapping is highly conventionalized in a language, it is more likely to systematically give rise to pragmatic interpretations in context, given that interpreters are assumed to be aware of the convention and its pragmatic implications. The morphosyntactic question forms investigated in Chapter 3 are an interesting case in point of how meaning con-

ventionalization might affect inferencing. As discussed in that chapter, it is unclear whether epistemic biases as discussed in the literature act as strong constraints on the interpretation of a question form, which assumes a stronger degree of meaning conventionalization, or whether, in combination with a question form, they can serve as pragmatic cues to intended meaning, which assumes a weaker degree of meaning conventionalization. In other words, while an epistemic bias might give rise to inferencing when a felicitous question form is used in context, as evidenced by the data reported in Chapter 3, it remains unclear how exactly the bias interacts with the form of a question and how strongly the mapping between form and meaning needs to be such that inferencing is triggered. Indeed, more research is needed on the conditions under which tracking a discourse bias – whether epistemic or otherwise – might be relevant for interpretation when processing linguistic forms with pragmatic meanings which are said to be sensitive to such biases.

The forms investigated in Chapter 5 further illustrate how meaning conventionalization can impact the derivation of a pragmatic inference. While the adjectives *actual*, *literal*, and *real* all have a conventional semantic meaning of their own, they also share a pragmatic meaning related to intensification. Differently from the case of questions with biases, in the case of *actual*, *literal*, and *real* there is a clearer divide between their truth-conditional (i.e. semantic) meaning and their non-truth conditional (i.e. pragmatic) meaning, despite both meanings alike having an epistemic flavor to them. What this means is that when interpreting an utterance containing an intensifying adjective the derivation of a pragmatic inference is not as dependent on tracking extra-sentential contextual information as in the case of questions with biases. Rather, it rests primarily on making sense out of the cues within the utterance itself, given the felicity conditions of the relevant pragmatic meaning. In more concrete terms, inferencing in such a case seems to involve weighing the contribution of each meaning associated with the adjectival form in light of other cues available in the sentence. As discussed in Chapter 5, in the case of a metaphoric statement like *He is a real weasel sometimes during games*, each of the different meanings associated with *real*

might give rive to a different interpretation when processing the adverb *sometimes* further downstream in the sentence, possibly resulting in strong enough a pressure such that an implicature is drawn, as attested in the data reported in the chapter.

Farther along a spectrum of meaning conventionalization are the forms investigated in Chapter 4. Indeed, unlike the English adjectives *actual*, *literal*, and *real*, whose two meaning components can occur – albeit with varying pragmatic import – in the exact same utterance context, the German adverbs *eigentlich* and *tatsächlich* have been pragmaticalized to the extent that they act as specialized pragmatic markers, their non-truth-conditional meaning being the most salient meaning component in their usage as discourse particles. What that means is that, unlike a lexical cue like an intensifying adjective or a morphosyntactic cue like a biased question form, a discourse particle is expected to be, by definition, a strong pragmatic cue to meaning, given its conventionalized status as a pragmatic marker. It therefore seems reasonable to conclude that the degree to which a given linguistic form – lexical or otherwise – is conventionally associated with a pragmatic meaning predicts the likelihood of an inference being drawn. In other words, forms which have more strongly conventionalized pragmatic meanings – in spite of any competing semantic or pragmatic meanings – are more likely to give rise to an implicature in context, this being potentially modulated by extra-sentential contextual factors like discourse biases which may license the usage of a particular form in the first place, as investigated in Chapter 3.

Another factor affecting the derivation of inferences on the basis of conventional meanings is the extent to which the relevant meaning is *indexically* associated with a particular linguistic form. Although indexicality is closely related to meaning conventionalizaton – both pertaining to the semiotic nature of the linguistic sign, it seems fruitful to treat indexicality, at least at a meta-analytic level, as a factor in its own right. As discussed in the introduction, a meaning is indexically linked to a form if that form indexes the relevant meaning through contextual co-occurrence. In other words, unlike in the case of form-meaning mappings where the relationship between the two sides of the linguistic sign is essentially non-indexical, in

the case of indexical form-meaning mappings this relationship is motivated by strength of distributional co-occurrence, that is, by how likely the form is to index its meaning in a given context. In the case of a form like a discourse or modal particle (see terminological discussion in Chapter 4), the nature of the indexical relationship is at its clearest: *eigentlich* and *tatsächlich* index referential contrast and agreement by virtue of occurring frequently in discourse contexts which provide evidential support to the conclusion that such is the relationship between the forms and their respective meanings. What is empirically interesting about such an indexical relationship is that it allows interpreters to reason pragmatically about the content of utterances, as evidenced by the data reported in Chapter 4. Particularly interesting is the fact that, as also shown in the data reported in Chapter 4, changes in the distributional properties of the mappings allow interpreters to infer that a form might not be a reliable indexical cue, as evidenced by the finding that unreliable usages of *eigentlich* and *tatsächlich* impact both the rate and the strength of the inferences drawn in context.

The indexicality of *eigentlich* and *tatsächlich* can be compared to that of *actual*, *literal*, and *real*. In the case of the adjectives, their pragmatic meaning of intensification is only indexically related to their form, whereas their lexical semantic meaning stands in more of a non-indexical relation to that same form. What this means is that, despite there being two meanings associated with the same adjectival form, one of them more strongly indexes a non-truth-conditional relation to the content of what the adjective modifies, meaning that it is interpreted as a marker of the subjective stance of the utterer in relation the propositional content of the utterance. The truth-conditional-meaning of 'actuality', 'literality', and 'reality' cannot be said to be indexically related to any element modified by *actual*, *literal*, and *real* the same way their subjective meaning of increased epistemic commitment can. This might, therefore, justify the inferential jump involved in the derivation of a conventional implicature evidenced in the processing of *He is a real weasel sometimes during games*, although, in this case, it is hard to disentangle the role of conventionalization from that of indexicality.

Where the relationship between indexicality, conventionalization, and inferencing is at its most intricate, and by extension where it is most difficult to disentangle one factor from the other, is in the case of question with biases. Indeed, the indexicality of complementary interrogative forms like *Didn't you ...?/ Did you not ...?* in English and *Hast du kein ...?/ Hast du nicht ein ...?* in German is said to be linked not only to particular evidential contexts but also to epistemic assumptions the interpreter must make with regard to the information state of the questioner. What the empirical picture suggests is that each alternative question form correlates with a different questioner stance, whereby, combined with evidential information which licenses the use of a negative question in the first place, each form carries a different meaning, one compatible with the straightforward semantic meaning of a negated polar question and the other signaling a particular expectation of what the answer to the question might be. Thus, in this case, pragmatic felicity seems to vary as a function of a complex mapping between different background assumptions onto alternative linguistic forms which, all else being equal, are equivalent in meaning. Given this complex relationship between specific extra-sentential discourse information and linguistic alternatives which are otherwise meaning-equivalent, one might argue that the indexical link between each question form and their meaning is not as tight as in the case of conventional pragmatic cues whose felicity is computed primarily against other cues in the same signal. This leads to a similar conclusion as above, namely that, from a processing perspective, inferencing on the basis of linguistic forms whose indexical link to a meaning might require strong evidential support from the extra-sentential context is less likely to occur compared to inferencing on the basis of forms whose indexical link to a meaning is made sufficiently salient by sentence-internal cues. Notice, however, how the exact nature of the indexical relation between a form and a meaning might give rise to different types of inferences in context. In the case of questions with biases, as just discussed, inferencing seems to involve tracking discourse information which biases the usage of particular question forms, which in turn might constrain a particular interpretation via a manner inference. In the case of the discourse particles discussed in

Chapter 4, on the other hand, the forms themselves are not a priori biased by any particular discourse context but rather, by virtue of being uttered, they might allow one to inference about the context via a conventional inference. In other words, while in the latter case the relationship between the forms and their meanings is strongly backward-looking, such that a particle indexes a discourse relation which might cue the interpreter to relate the content of the utterance to previous information in the discourse context, in the former case the relationship between the forms and their meanings seems to be only indirectly backward-looking, such that a question form indexes a questioner stance which in light of additional assumptions might cue the interpreter to reason about the questioner's expectations regarding the content of the utterance.

In conclusion, a general theory of pragmatic interpretation must, therefore, account for the role of indexicality alongside that of meaning conventionalization. While meaning conventionalization can, at least in theory, be stipulated and even empirically characterized independently from any contextual cues, indexicality, by definition, requires understanding the exact contextual factors which constrain the co-occurrence of a form and a particular meaning. As discussed here, strength of interpretation can be construed as varying as a function of the accumulated evidential support for a particular interpretation, which may include both truth-conditional and non-truth-conditional meaning components and which can be modulated by the degree to which these different meaning components are conventionalized with particular linguistic forms. An additional assumption is that the evidential weight of any given form in context is influenced by whether that form is indexically related to its relevant meaning, unlike forms which have essentially non-indexical relations with their meanings and are thus, pragmatically, much less context-dependent. This relationship can be captured in a theoretical space as illustrated in Figure 6.2, which is extension of Fig. 6.1 above. As suggested by both Fig. 6.1 and 6.2, the relationship between pragmatic import and non-truth-conditionality is not expected to be linear, such that, on part of the space, forms that contribute meanings which are less truth-conditional are expected to have

203

a stronger relative impact on interpretation compared to forms that contribute meanings which are more truth-conditional, while, on another part of the space, the reverse is true. However, as suggested by Fig. 6.2, this relationship is mediated by the indexicality of a form-meaning mapping, such that forms that have stronger indexical links to their meanings (e.g. modal particles) are expected to have a stronger relative impact on interpretation compared to forms that have weaker indexical links to their meanings (e.g. intensifying adjectives).
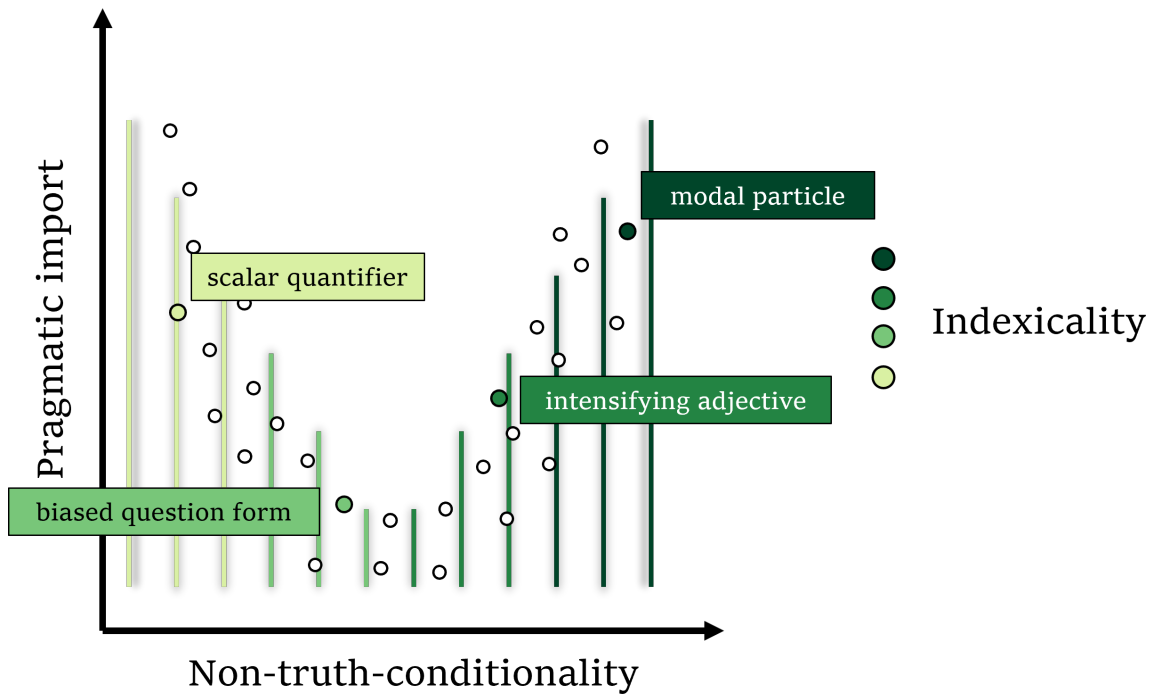
Figure 6.2 Hypothesized relationship between the pragmatic import of a given linguistic form and the nature of its meaning contribution in context as modulated by the indexicality of the form-meaning mapping. Each point represents one potential form/ phenomenon in such a theoretical space. The colored points represent the approximated location of the phenomena investigated in this dissertation. The vertical bars represent varying degrees of indexicality, where darker bars represent more indexical areas of the space.

6.3 Predictability and prediction in pragmatic processing

So far, I have discussed a number of factors which may impact the likelihood of drawing different types of inference in context. On the basis of that discussion, I have argued for a view of pragmatic inferencing as a graded interpretation process where interpretation can be equated with abductive reasoning guided by evidential support in context. As the findings reported in the different studies of this thesis show, whether or not interpreters actually draw an inference depends both on factors argued to affect interpretation according to standard assumptions in pragmatic theories as well as factors which have not received much attention in (experimental) pragmatic research. In what follows, I discuss in more detail how a supposed process of evidence accumulation can unfold incrementally and how it can empirically affect online pragmatic processing as measured in experimental psycholinguistic tasks.

Common to all four empirical studies in this thesis was the concept of predictability, which was defined in the introduction as the likelihood of a particular cue being predicted or anticipated in light of previous cues. While predictability is a general feature of any information processing system (Shannon, 1948), in the present thesis I made the assumption that pragmatic processing is – or at least relies on – such an information processing system. Another central feature of information processing systems is that they operate on uncertainty, whereby the system is geared towards minimizing prediction error and it therefore relies on conditional or relative predictability (Shannon, 1948). Along these general lines, language comprehension, and in particular real-time comprehension, has been equated with an expectation-managing mechanism which processes cues and generates predictions for subsequent cues given the accumulated evidence found in already processed material (e.g. Levy, 2008). Therefore, in such a framework, the predictability of any given form in a linguistic signal can be linked to its (conditional) likelihood in a concrete context, which, depending on the phenomenon at hand, can be understood either as the sentential context in the narrow

sense or the larger discourse context a particular signal is embedded in.

In chapters 2, 3, and 5 of this thesis, which employed a word-by-word self-paced reading paradigm, the likelihood of expecting any given word in an unfolding sentence was said to be linearly related to the time taken to read that word, as measured in the experimental task. Crucially, this means that words that were read more slowly were assumed to be harder to parse and/ or integrate, which in turn assumes that they were less likely sentence/ discourse continuations. In Chapter 4 of this thesis, which employed a mouse-tracking task, the likelihood of expecting one of two possible referents as a sentence continuation was said to be linearly related to the mouse trajectories measured within a particular time window in the experimental task. Rather than just listening for comprehension, the task involved moving the mouse towards the referent most likely to be compatible with the experimental discourse. Crucially, this means that trajectories that shifted earlier towards the correct referent without subsequent change of course were assumed to index a higher likelihood of expecting that referent compared to trajectories which shifted later towards the same referent.

To recapitulate the main findings of the different empirical chapters, in chapters 2 and 3 I did not find evidence that comprehenders predicted incoming linguistic material after processing the relevant pragmatic cues, at least not as predicted by my working hypotheses. In chapters 4 and 5, on the other hand, I did find evidence that comprehenders predicted incoming linguistic material after processing the relevant pragmatic cues, as predicted by the accounts entertained in each of those chapters. But what exactly does it mean – both empirically and theoretically – to state that comprehenders did or did not predict incoming material in a sentence?

Empirically speaking, the notion of prediction relates to three separate factors in an experimental processing task: what is measured in the task, what is theoretically expected to be predicted in the stimuli, and what the linking hypothesis between the empirical measurements and the theoretical constructs of interest might be. The two methodologies used in

this thesis, self-paced reading and mouse-tracking, provide different proxies for prediction in online processing. While tracking mouse trajectories as participants move their cursor towards different referents on a screen allows measuring a close approximation of real-time anticipation, recording reading times as participants move from word to word in a sentence allows approximating prediction in terms of conditional expectancy, which in this particular case was grounded in the notion of surprisal (see Introduction and Chapter 2 for a detailed discussion). That is to say that shifts in a mouse trajectory either towards or away from a referent recorded before the onset of a critical noun index true anticipation of that noun within the constraints of the experimental task, whereas reading times recorded at a critical word index the difficulty in processing and integrating that word into an unfolding sentence/ discourse model, which does not necessarily involve anticipating the word in a strict sense but rather computing its predictability on the spot as a function of surprisal.

The exact role different forms of linguistic prediction might play in processing – whether strict anticipation/ pre-activation of linguistic material or expectation by predictability – has been discussed in detail elsewhere in the psycholinguistic literature concerned with predictive processing (e.g. Huettig, 2015; Kutas et al., 2011). What matters, principally, for the present discussion is that the same prediction mechanism is assumed to underlie different forms of pragmatic processing regardless of whether evidence for prediction is found in mouse trajectories or reading times, despite each of the two methods indexing prediction in a qualitatively different way. Indeed, core to the account of prediction discussed here is the notion of expectation: participants move their mouse towards a referent because that referent is assumed to be a more expectable continuation to the sentence they are hearing, similarly to how participants read a word faster because that word is assumed to be a more expectable (i.e. less surprising) continuation to the sentence they are reading. In both cases, expecting a particular word is taken to mean that comprehenders have generated predictions about the likelihood of that word serving as a pragmatically felicitous continuation to a sentence, as per an expectancy/ constraint-based approach to pragmatic processing (Degen and Tanenhaus,

2019; Rohde and Kurumada, 2018).

When it comes to the possible reasons as to why comprehenders predict linguistic material based on pragmatic reasoning, aside from the theoretical assumptions linked to each particular type of pragmatic inference, there is the assumption that the forms investigated in each study are pragmatically relevant for interpretation as they can, at least in principle, provide enough evidential support to trigger an inference. As discussed above, processing signatures recorded in an experimental task can provide empirical evidence of that inferential process taking place. As such, while the findings from the two chapters where evidence for prediction is found elucidate what the inferential process might look like as per the theoretically-motivated assumptions, the cases where evidence for an inference is not found are perhaps even more elucidating with regard to any potential underlying principles. In Chapter 2, contrary to what was predicted by the study hypotheses, comprehenders did not predict the shape terms in referring expressions of the sort *Alle/ Einige Dreiecke auf dem Bild sind orange* when reading those expressions after looking at a picture depicting two sets of geometric shapes. Crucially, participants were expected to predict the shape term when the picture biased a particular referential mapping, as when only one of the shape arrays was homogeneous in color. The reading times measured in this condition were compared to the reading times measured in a condition in which both shape arrays were either homogeneous or heterogeneous in color, where participants were not expected to predict any particular shape term. While a null result would correspond to a lack of difference in reading times between the two experimental conditions, what was actually found was facilitation at the shape term in the unbiasing condition. This raises a number of interesting questions, discussed in detail elsewhere in this thesis. Crucial for the current discussion is the fact that the findings do not constitute null evidence for prediction by informativity but rather they provide evidence for prediction of a different sort, namely one which is guided by a pressure to integrate crucial semantic information as quickly as possible. This suggests that different pressures are at play during incremental processing, and that one pressure may override the

other, even when the overridden pressure is theoretically assumed to be of central importance to the pragmatic interpretation of a given sentence. Such cases provide an important empirical insight, namely that actual processing – and by extension actual inferencing – is more intricate than standard pragmatic theories assume interpretation to be. Revising theoretical assumptions on the basis of challenging empirical data and designing empirical tests which can reliably expand the limits of existing theoretical frameworks remains a challenge for both present and future (experimental) pragmatic research.

Chapter 3 provides another such interesting empirical insight, one which should nevertheless be treated carefully when drawing theoretical conclusions given the restricted empirical picture available on the topic. In Chapter 3, contrary to what was predicted by the study hypotheses, comprehenders did not predict the semantic cues immediately following the question forms in questions of the sort *Didn't you buy a ticket?/ Did you not buy a ticket?*. And yet, at least in the English data, there was strong evidence for prediction further downstream in the signal, at the last semantic cue in the sentence. Even though the predictions derived from the theoretical account were not stringent when it comes to the exact locus of prediction in the sentence, one can conclude on the basis of the results that the assumption that an epistemic bias combined with a felicitous question form can serve as a pragmatic constraint on interpretation might need to be relaxed, such that comprehenders seem to require further semantic information in order to be able to reason pragmatically about the content of the utterance and thus to predict subsequent material in the sentence. All in all, while this suggests that prediction is in principle possible under a relaxed assumption of pragmatic felicity operating as a constraint on interpretation, further fleshing out the conditions under which this assumption might hold seems like a necessary next step in advancing a pragmatic theory of question bias interpretation, and more generally of question interpretation, an underinvestigated topic in psycholinguistics and experimental pragmatics.

Ultimately, what the studies discussed here have shown is that pragmatic processing seems to unfold based on constraints which are, on the one hand, more varied and context-

dependent as is usually assumed in the literature. On the other hand, these different constraints all seem to share a common explanatory basis, namely the assumption that inferencing and predictive processing advance based on the accumulation of evidential support for a particular interpretation, at least in the case of phenomena which can be canonically treated as implicatures. This seems to be in line with comprehension models proposed in the psycholinguistic literature which feature both prediction and integration as central aspects of incremental language processing (e.g., Ferreira and Chantavarin, 2018; Kuperberg and Jaeger, 2016). Importantly, these models assume that information from lower levels of representation can actively contribute to the build-up of evidential support at higher levels of representation, such that richer representations are more likely to give rise to (strong) predictions about subsequent information, whether that information relates to the meaning of an utterance or the form of particular linguistic cues. Future research focused on pragmatic prediction should address more explicitly the interconnections between different types of pragmatic inferences, attempting to answer some outstanding foundational questions such as: Is degree of context-dependency the primary underlying factor distinguishing one type of pragmatic inference from another? Can a graded view of interpretation be reduced to a process of evidence accumulation or is there a role to be played by complex perspective-taking? Does inferencing depend primarily on the computation of overt linguistic commitments or does it rely on some form of intention recognition?

6.4 Epistemicity and inferencing

The studies discussed in chapters 3, 4, and 5 of this thesis all relate to the interpretation of so-called markers of epistemicity. Epistemicity was defined in the introduction as the subdomain of linguistic modality concerned with both the marking of information source and the marking of language users' personal commitment to the truth, plausibility, or relevance of an utterance. While epistemicity can be expressed linguistically via grammatical as well

as lexical means, different languages make available to their users different types of devices. Crucial for the present discussion is the fact that such devices, as diverse as they are, commonly express meanings which blur the line between what is prototypically considered to be the marking of personal feeling/ attitude and the marking of status of knowledge (Gray and Biber, 2012).

In this thesis I investigated two lexical devices and one morphosyntactic device across nine different forms in two distinct languages. Despite the meanings associated with these different devices being quite different from one another, all three devices alike have an epistemic flavor to what they express, one which might nonetheless not necessarily be transparent in their respective pragmatic functions. As has been discussed in section 6.1, the degree to which these different meanings are conventionalized with their respective linguistic forms is a key difference between the three devices analyzed in this thesis. As a matter of fact, meaning conventionalization seems to be a key factor impacting the likelihood of each device giving rise to an inference in context, regardless of whether they mark epistemic confidence in the truth of a proposition (question with biases), a subjective commitment to the truth of a proposition (intensifying adjectives), or a procedural stance relative to a proposition (modal particles).

In order to generate higher-order predictions about the likelihood of different epistemicity markers giving rise to a pragmatic interpretation, one can place them on a continuum according to the extent to which their relevant pragmatic meaning is claimed to be conventionalized with their form. This can be achieved by stipulating an approximate degree of conventionalization based on accounts of their function and usage profiles, ideally triangulated from both theoretical and empirical accounts as well as from both qualitative and quantitative sources. For instance, based on what is known about their usage profile, the discourse particles *eigentlich* and *tatsächlich* can be placed at the upper end of the continuum, given their function as fully conventionalized pragmatic markers which can be considered to be a more specialized type of discourse particle, what is known in the literature as a modal particle.

The intensifying adjectives *actual*, *literal*, and *real* might lie somewhere in the middle of the continuum, whereby their pragmatic meaning is conventionally associated with their forms but they are not considered to be specialized pragmatic markers, such that they still express their conventional semantic meaning in the same context of utterance. Finally, *Didn't you ...?/ Did you not ...?* and *Hast du kein ...?/ Hast du nicht ein ...?* might lie at the lower end of the continuum, whereby their pragmatic meaning seems to be less conventionalized than their conventional semantic meaning, its felicity being constrained by extra-sentential factors which might make it salient enough in context such that it potentially becomes pragmatically relevant.

What this graded view of meaning conventionalization might mean for a theoretical account of inferencing is that the less conventionalized a given pragmatic meaning is, the less likely it is to trigger an inferential process in the first place, as discussed earlier. Crucially, this provides a general principle of pragmatic interpretation which extends beyond individual meanings and any idiosyncratic pragmatic assumptions and at the same time captures the natural gradience in meaning contribution characteristic of different linguisic devices, as tentatively represented in Fig. 6.2. Needless to say, meaning conventionalization is entwined with additional factors – it itself can vary as a function of the linguistic nature of the particular device under consideration (e.g. whether a lexical or a (morpho)syntactic device), the potential polysemous character of a given device, and by extension the nature of the other meanings which might compete with the relevant pragmatic meaning of a device (i.e. whether they are also pragmatic or semantic in nature). Still, conceptualizing variation in pragmatic import in terms of variation in the conventionalization of a given pragmatic meaning might be a fruitful way of stipulating the weight any given pragmatic cue might play in online processing. Not only that, it can also be a helpful principle in empirically characterizing pragmatic import, whether in experimental laboratory investigations or via other methods in linguistic and cognitive scientific research.

Ultimately, epistemicity markers seem to be a particularly rich and elucidating testing

ground for pragmatic (processing) theories, given that such markers function, by definition, as overt markers of stance and evidence accumulation. Future research should therefore aim to disentangle the pragmatic potential of different types of epistemicity markers as a function of their linguistic status and degree of pragmaticalization. Perhaps more importantly, it should aim to investigate the converging meaning contribution of multiple cues both within a signal and across signals, especially in cases where different cues might signal conflicting evidence and where meaning might be negotiated by interlocutors piecemeal. All in all, understanding how both linguistic and non-linguistic cues can provide converging evidence for particular interpretive conclusions, especially at a level that extends beyond individual signals, seems like a particularly exciting new avenue for cognitively-oriented pragmatic research. The core principle of interpretation as a process of incremental accumulation of evidence extends beyond analyses situated at the signal level and affects understanding in ways which also go beyond what a processing-level approach alone might provide in terms of explanatory power. Future research in cognitively-oriented pragmatics should therefore aim to bridge cognitive processing accounts with those that understand pragmatic language use as an iterative process of forming and ascribing actions in and through communication, where notions like indexicality play a crucial role. After all, pragmatics is about meaning and interpretation in context and, as such, by its very definition, it should lie at the nexus between producer and comprehender, action and response.

REFERENCES

Aparicio, Helena, Christopher Kennedy, and Ming Xiang (2018). "Perceived Informativity and Referential Effects of Contrast in Adjectivally Modified NPs". In:

Ariel, Mira (2016). "Revisiting the typology of pragmatic interpretations". In: *Intercultural Pragmatics* 13.1, pp. 1–35.

Degen, Judith and Michael K Tanenhaus (2019). "Constraint-based pragmatic processing". In: *The Oxford handbook of experimental semantics and pragmatics*, pp. 21–38.

Ferreira, Fernanda and Suphasiree Chantavarin (2018). "Integration and prediction in language processing: A synthesis of old and new". In: *Current directions in psychological science* 27.6, pp. 443–448.

Gotzner, Nicole and Jacopo Romoli (2022). "Meaning and Alternatives". In: *Annual Review of Linguistics* 8.1, pp. 213–234.

Gotzner, Nicole, Stephanie Solt, and Anton Benz (2018). "Adjectival scales and three types of implicature". In: *Semantics and Linguistic Theory*. Vol. 28, pp. 409–432.

Gray, Bethany and Douglas Biber (2012). "Current conceptions of stance". In: *Stance and voice in written academic genres*. Springer, pp. 15–33.

Horn, Laurence R (1985). "Metalinguistic negation and pragmatic ambiguity". In: *Language*, pp. 121–174.

Huettig, Falk (2015). "Four central questions about prediction in language processing". In: *Brain research* 1626, pp. 118–135.

Kaiser, Elsi and John C Trueswell (2004). "The role of discourse context in the processing of a flexible word-order language". In: *Cognition* 94.2, pp. 113–147.

Kuperberg, Gina R and T Florian Jaeger (2016). "What do we mean by prediction in language comprehension?" In: *Language, cognition and neuroscience* 31.1, pp. 32–59.

Kurumada, Chigusa et al. (2014). "Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings". In: *Cognition* 133.2, pp. 335–342.

Kutas, Marta, Katherine A DeLong, and Nathaniel J Smith (2011). "A look around at what lies ahead: Prediction and predictability in language processing." In:

Levinson, Stephen C (2016). "Turn-taking in human communication–origins and implications for language processing". In: *Trends in cognitive sciences* 20.1, pp. 6–14.

Levy, Roger (2008). "Expectation-based syntactic comprehension". In: *Cognition* 106.3, pp. 1126–1177.

Macagno, Fabrizio (2012). "Presumptive reasoning in interpretation. Implicatures and conflicts of presumptions". In: *Argumentation* 26.2, pp. 233–265.

Rohde, Hannah and Chigusa Kurumada (2018). "Alternatives and inferences in the communication of meaning". In: *Psychology of Learning and Motivation*. Vol. 68. Elsevier, pp. 215–261.

Shannon, Claude Elwood (1948). "A mathematical theory of communication". In: *The Bell system technical journal* 27.3, pp. 379–423.

Stivers, Tanya et al. (2009). "Universals and cultural variation in turn-taking in conversation". In: *Proceedings of the National Academy of Sciences* 106.26, pp. 10587–10592.

Tourtouri, Elli N, Francesca Delogu, and Matthew W Crocker (2021). "Rational Redundancy in Referring Expressions: Evidence from Event-related Potentials". In: *Cognitive Science* 45.12, e13071.

Wu, Shiyu and Zheng Ma (2020). "How is Chinese reading affected by under-specification and over-specification? Evidence from self-paced reading experiments". In: *Journal of Pragmatics* 155, pp. 213–233.