Challenges for a Computational Explanation of Flexible Linguistic Inference

Marieke Woensdregt (marieke.woensdregt@ru.nl)
Language and Computation in Neural Systems, Max Planck Institute for Psycholinguistics, The Netherlands
Department of Cognitive Science and Artificial Intelligence, Radboud University, The Netherlands
Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands

Mark Blokpoel (m.blokpoel@donders.ru.nl) **Iris van Rooij (i.vanrooij@donders.ru.nl)**
Department of Cognitive Science and Artificial Intelligence, Radboud University, The Netherlands
Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands

Andrea E. Martin (andrea.martin@mpi.nl)
Language and Computation in Neural Systems, Max Planck Institute for Psycholinguistics, The Netherlands
Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands

Please cite as: Woensdregt, M., Blokpoel, M., van Rooij, I., & Martin, A. (2024, July). Challenges for a Computational Explanation of Flexible Linguistic Inference. Paper published at MathPsych / ICCM 2024. Via mathpsych.org/presentation/1522.

Abstract

We identify theoretical challenges for developing a computational explanation of flexible linguistic inference. Specifically, the human ability to interpret a novel expression (like *mask-shaming*), where inferring plausible meanings requires integrating relevant background knowledge (e.g., COVID-19 pandemic). We lay out (i) the core properties of the phenomenon that together make up our construal of the explanandum, (ii) explanatory desiderata to help make sure a theory explains the explanandum, and (iii) cognitive constraints to ensure a theory can plausibly be realised by human cognition and the brain. By doing so, we reveal the 'force field' that theories of this explanandum have to navigate, and we give examples of tensions that arise between different elements of this force field. This is an important step in theory-development because it allows researchers who aim to solve one part of the puzzle of flexible linguistic inference to keep in clear view the other parts.

Keywords: language comprehension; inference; theory-development; computational explanation; meta-theory

Introduction

Language use is remarkably flexible. One aspect of this is that humans appear to be able to integrate different kinds of knowledge in novel ways when interpreting utterances. In this paper, we focus specifically on humans' ability to come up with possible interpretations of neologisms, such as *mask-shaming*. Coming up with a plausible interpretation of such a novel expression arguably requires an ability to relate knowledge of the meaning of the words and how they are combined, to broader contextual or world knowledge (e.g., about the COVID-19 pandemic; see Figure 1). Explaining this phenomenon raises several theoretical challenges[1]: What *is* the phenomenon really? What counts as a *good* explanation? The aim of this paper is to outline those challenges. Importantly, we consider how these challenges will interact, which brings into clear view a 'force field' that explanations of flexible linguistic inference need to navigate.

The contribution we make in this paper takes inspiration from several sources. First, Adolfi, van de Braak, and Woensdregt (2023) argue that theoretical problem-finding (as opposed to empirical problem-solving) is an important scientific contribution in its own right. This activity involves not just characterising the phenomenon, but also identifying the theoretical constraints that determine what makes a good explanation. Second, Guest (2024) and Guest and Martin (2023) argue that as scientific practitioners, we can make meta-theoretical commitments about criteria that make a theory good. Guest (2024) calls upon scientists to characterize and examine the criteria we use to adjudicate over theories by building and sharing what Guest and Martin (2023) and Guest (2024) dubbed a *metatheoretical calculus*: a formal system that describes the process by which theories are evaluated and pitted against each other in a particular (sub)field. Finally, Blokpoel (2018) argues that developing a computational-level model (i.e., a formalised theory) of a cognitive capacity is like sculpting. The scientist has to start out with a sufficiently large block of material (i.e., model/theory) that can capture the entire capacity (i.e., is *generatively sufficient*), and can then figure out which parts to chisel away by applying various *computational-level constraints* (e.g., tractability).

In this paper, we take inspiration from these approaches, and apply them specifically to the phenomenon of flexible linguistic inference. That is, the human ability to flexibly interpret neologisms upon first encounter, in a way that appears to require integrating linguistic knowledge with world knowledge. We start by outlining the specific phenomenon in language comprehension that we want to explain, in the form of three key properties, in Section *The explanandum*. Next, inspired by Guest (2024), Guest and Martin (2023), Blokpoel (2018), and Adolfi, van de Braak, and Woensdregt (2023), we discuss two classes of constraints (*Constraints on the explanans)* that we deem particularly relevant for theories of

[1] We would like to preempt the possible presupposition that large language models (LLMs) would already address these challenges. LLMs do not provide any precise characterisation of the explanandum (human flexible linguistic inference), nor are they explanatory (Guest & Martin, 2023; Bender & Koller, 2020; van Rooij et al., 2023; van Rooij, 2022).

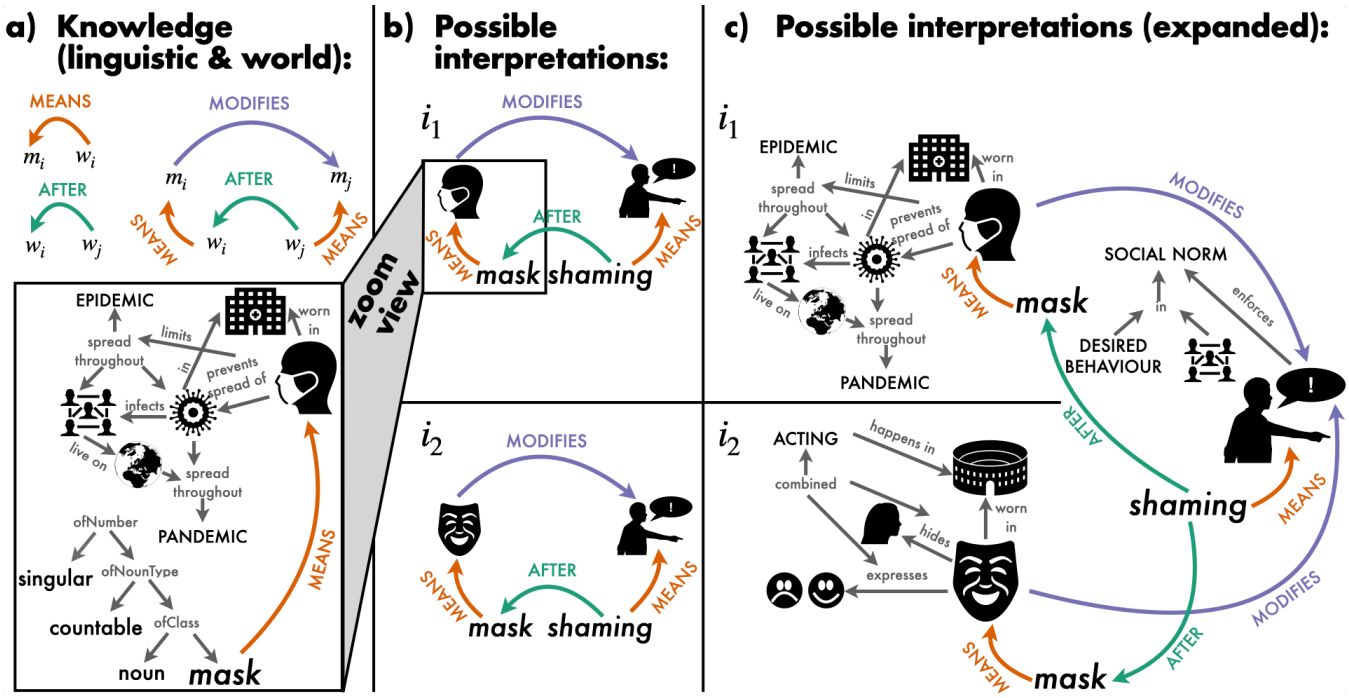

Figure 1: Illustrative example of our construal of the explanandum: The ability to come up with a plausible interpretation of the neologism *mask-shaming*. From left to right: (a) structured representations of stored knowledge (including grammatical, semantic, and world knowledge) are involved in building (b) structured representations of possible interpretations (here, two possible interpretations, i_1 and i_2 are shown, with a zoomed-in view of i_1 to illustrate the further structured knowledge that is associated with these abstract representations). Finally, (c) given the background knowledge associated with the semantic representations, and assuming this word is interpreted within the context of the COVID-19 pandemic, i_1 is more plausible.

this explanandum. First, in Section *Explanatory desiderata*, we discuss two metatheoretical commitments that can help make sure a given theory really explains the explanandum of interest. Second, in Section *Cognitive constraints*, we discuss two metatheoretical commitments that can help make sure the theory can also be plausibly realised by human cognition and the brain. Finally, in Section *Challenges for explaining flexible linguistic inference*, we highlight some examples of tensions that may arise between these properties and constraints.

The properties of the phenomenon and constraints on the explanans that we highlight in this paper are not exhaustive; we see them as necessary but possibly not sufficient. However, by outlining the explanandum, several constraints on the explanans, and some of the tensions that can arise between these, we shed light on the force field that theories of flexible linguistic inference need to navigate. This provides a foundation from which further theory-development can depart.

The explanandum

In this section, we describe the core properties of the phenomenon we want to explain; that is: our construal of the *explanandum*. It appears to be the case that humans, under the right circumstances (given a shared language, shared world knowledge, and shared motivation to achieve mutual understanding) are able to interpret novel expressions in a way that requires knowledge not just of the word meanings and grammar of the language, but also broader contextual or world knowledge. And that these different kinds of knowledge are flexibly integrated in this process of meaning inference. For example, the first time you heard the term *mask-shaming*, you were probably able to come up with a sensible (not necessarily correct) interpretation of what this might mean, in the context of the COVID-19 pandemic (Blokpoel, Wareham, Haselager, Toni, & van Rooij, 2019). Figure 1 shows a possible construal of what such an inference process might involve. Below, we discuss three properties that we believe together form the core of this explanandum. Our construal leaves out other components of language comprehension that also require explanation, such as segmentation (i.e., turning a continuous stream of sound or sign into discrete units; Adolfi, Wareham, & van Rooij, 2023) and word recognition (i.e., mapping a sequence of phonemes onto a lexical representation; Lahiri & Marslen-Wilson, 1991; McQueen, 2007). These capacities are outside the scope of this paper, as our construal of the explanandum does not rely on any particular theory of them.

Language comprehension is compositional

To understand the meaning of a linguistic expression (a phrase or sentence), one doesn't need to have come across it

as a whole before. Instead, we can most often infer the meaning of the whole by knowing the meanings of the parts (lexical semantics) and how the structure of the whole influences its meaning (syntax). The fact that (most often) the meaning of the whole is a function of the meanings of the parts and the way in which those are combined, makes natural languages *compositional* (Martin & Baggio, 2020; Partee, 1995; Pylkkänen, 2020). This compositionality buys us a high degree of systematicity and productivity (i.e., we can produce and understand utterances we have never come across before) (Szabó, 2004; Martin & Baggio, 2020; Pylkkänen, 2020).

Language comprehension requires building abstract hierarchical structure from linearly incoming sensory input, on the fly (Hagoort, 2019). Martin (2016, 2020) captures this computationally as a process of perceptual inference, in which incoming sensory cues are transformed into increasingly abstract structures through activation of stored knowledge representations. This computational model can account for cases of language comprehension in which the compositional meaning can be inferred directly from the stored language knowledge and its mapping to conceptual knowledge. However, humans are also able to infer the possible meanings of novel expressions in a way where semantics, syntax and compositionality alone are not enough.

Language comprehension involves world knowledge

Knowledge of the meanings of words (lexical semantics) is often not independent from world knowledge.[2] Hagoort et al. (2004) showed that in language comprehension, general world knowledge is integrated simultaneously with lexico-semantic knowledge (see also Hagoort & van Berkum, 2007). Using EEG, they showed that the event-related potential (ERP) component associated with semantic integration (the N400) looks similar in terms of timing, shape, and location when reading sentences like "the Dutch trains are white and very crowded" (a violation of world knowledge for the Dutch participants, who know that Dutch trains are yellow) compared to "the Dutch trains are sour and very crowded" (a semantic violation, because the semantic features of the predicate "sour" do not fit those of its argument "trains"). This is empirical evidence against the classic two-step model of language interpretation in which first the 'local' meaning of the compound expression is computed, and world knowledge is only integrated in a second step, to work out what the expression really means. Instead, Hagoort and van Berkum (2007) show that world knowledge is brought to bear on utterance interpretation as soon as it's available (Just & Carpenter, 1980; Hagoort & van Berkum, 2007; Hagoort, 2019).[3]

[2]To illustrate how word meanings are underdetermined in the absence of world knowledge, Hagoort, Hald, Bastiaansen, and Petersson (2004) provide the following example: The word "finish" means something different in the phrase "Mary finished the book" (which implies she completed reading or writing it) compared to "the goat finished the book" (which implies the goat ate or destroyed it).

[3]For a computational model of this integration of world knowledge during incremental comprehension, see Venhuizen, Crocker, and Brouwer (2019).

The importance of world knowledge for language comprehension becomes especially apparent when interpreting novel expressions such as *mask-shaming* (see Figure 1). We posit that in addition to building compositional structure based on stored and structured language and world knowledge, this requires inferring new relationships between the incoming sensory cues and (potentially novel) conceptual representations. This may involve *abductive inference*, where novel candidate hypotheses to explain a given observation are generated (in this case: possible interpretations of a novel linguistic expression) (Blokpoel et al., 2019). Explaining this ability may require a computational model that can reach across different capacities in cognition and capture systematicity between structured representations of incoming language input and structured representations of world knowledge.

Language comprehension is incremental

Words (or signs) come in incrementally during language comprehension, in linear order (although signed languages allow for more simultaneity than spoken languages; Slonimska, Özyürek, & Capirci, 2020). Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy (1995) showed that linguistic utterances are also *processed* incrementally, not just syntactically but also semantically and in context. They showed that participants seek to establish reference in context immediately, as soon as words come in. More recently, Hagoort (2019) reviewed various psycholinguistic studies on meaning-making, and concluded that complex meaning is created on the fly, through a unification operation that takes lexical meanings and context as its input and outputs a situation model (see Zwaan & Radvansky, 1998, for more on situation models).

Sedivy (2007) reviews the psycholinguistic literature on incremental language processing in the context of theories of pragmatic inference (i.e., going beyond the literal meaning of an utterance to figure out what it means in context). She shows that there is at least some evidence from self-paced reading experiments that participants rapidly integrate expectations based on the informativeness of different possible referring expressions given the context. This suggests that also the pragmatic integration of context happens incrementally.

The combination of incremental and immediate language processing means that as hierarchical representations are being built from a linear input sequence, the set of possible interpretations and their hierarchical structure may change and need to be revised as new words come in. This means that a computational-level theory of flexible linguistic inference needs to be able to produce intermediate output when provided with only partial input sequences.

Constraints on the explanans

In this section, we highlight two classes of metatheoretical constraints or commitments that we deem particularly important for the explanandum described above. These two classes of constraints are somewhat different in nature: The *Explanatory desiderata* have to do with whether a given theory can really explain the phenomenon, and the *Cognitive*

constraints have to do with whether the theory can be plausibly realised by human cognition and the brain. Blokpoel (2018) argues that developing a computational-level model (i.e., a formalised theory) of a cognitive capacity is like stone carving a sculpture out of a block of marble. First, the modeller needs to make sure the block of marble they start with (the 'starting theory') is large enough to capture the entire explanandum (i.e., generatively sufficient). Otherwise, they would have to glue parts back on later, which, in this analogy, corresponds to adding ad-hoc components to the model. Second, they can start chiseling down the sculpture based on various constraints, until the model provides a precise fit of the cognitive capacity (i.e., the explanandum). In this paper, we build on this metaphor: We view the explanatory desidarata described below as characteristics of the block of stone that the sculptor starts out with, and the cognitive constraints as informing the chiseling process. This sculpting analogy also allows us to illustrate what consequences it has for the later chiseling process if the explanatory desiderata are violated.

Explanatory desiderata

Below, we discuss two metatheoretical desiderata that we consider important for a theory of a cognitive process to be explanatory. We also discuss the consequences if these desiderata are not satisfied: such a theory is likely to break apart during further chiseling based on cognitive constraints (Figure 2). This highlights the importance of having 'good quality material' to start with: A theory that (i) does not assume what it's trying to explain, and (ii) is not piecemeal.

Explaining without assuming Explanations of cognitive processes can be described on the computational level as a function that maps from input to output. That is, we can formalise a hypothesis about *what* a given cognitive capacity does (i.e., a computational-level explanation), as a function $f : I_f \rightarrow O_f$ that specifies for each input $i \in I_f$ its corresponding output $o \in O_f$ (Marr, 1982). Such a computational-level theory constrains the set of possible algorithmic-level and implementation-level specifications that are consistent with it (Blokpoel, 2018). By explaining without assuming, we mean that on the computational level, the theoretician should not slip by assuming that that which is to be explained is part of the input. Instead, the theory has to explain how a given property of the explanandum is part of the output *as a function of* the input. If, instead, this property that is in need of explanation is assumed, without explaining how it arises or where it comes from, the theory can be considered 'hollow', and this may reveal itself upon later chiseling.

Let us take the compositional nature of language comprehension as an example. The input in this case should be a linear sequence of words, and the output should be a hierarchical representation of the compositional structure that arises from the interaction between the meanings of the words and the way in which they are combined. If compositional structure is already present in the input to this function, it is assumed, rather than explained. If, instead, the formalisation of the model provides a specification of the output *as a function of* the input, where (some of) the output has compositional structure but the input does not, we can state that it explains compositionality without assuming it. Note that this definition of explaining without assuming is independent of the specific definition of compositionality one is working with.

Non-piecemeal We consider a theory piecemeal if it makes use of different components (e.g., several separate computational processes) to explain different aspects of the explanandum of interest. The worry with such a piecemeal explanation is that it also requires an explanation of how these different components (e.g., computational processes) interface. This process of 'glueing' the different components back together may turn out hard (especially if these different component explanations were developed independently from one another), for example because they have incompatible assumptions. There can be valid reasons to conclude that a piecemeal explanation, postulating several different computational processes, is in fact necessary. However, aiming for a non-piecemeal approach first, can potentially avoid the hard problem of having to glue parts back together later. Furthermore, by adopting such a non-piecemeal approach, the limits of reaching such a unified, non-piecemeal explanation for a given explanandum will eventually be discovered, if they exist. This does require starting out with a well-specified and clearly carved out explanandum.

Let us take the different levels of organisation we find in linguistic expressions (from phonemes to morphemes to words to phrases to sentences) as an example. If our explanation entails a computational process that could be applied iteratively to build up interpretations from the smallest meaningful linguistic unit (morphemes) up to entire sentences, it can be considered non-piecemeal. If, instead, it has to postulate multiple computational processes in order to account for different levels of linguistic analysis, it is more piecemeal in nature. See Martin (2020) for an example of a non-piecemeal approach to explaining language comprehension.

Cognitive constraints

Here, we discuss two meta-theoretical constraints that are specific to theories that aim to explain cognitive capacities. These two constraints are necessary (i.e., any explanation that doesn't satisfy these two constraints will inherently not explain the phenomenon), but not sufficient (i.e., other constraints that we do not discuss may also apply, meaning any explanation that does satisfy the two constraints will not automatically provide a good explanation of the phenomenon). For example, given the explanandum of flexible linguistic inference, one may further want to ensure that the theory is consistent with insights from psycholinguistic research.

Computational tractability Human minds are resource-bounded. That is, we have limited time and memory resources. This means that human minds (just like any other resource-bounded system, such as a computer) can only com-

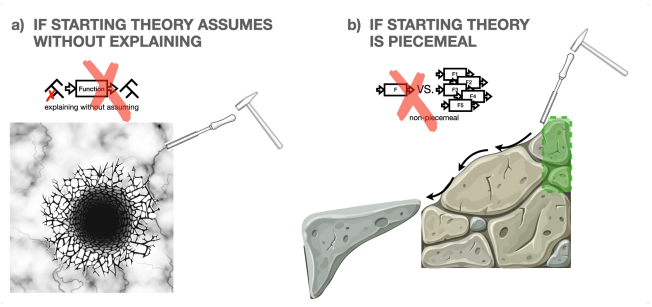

Figure 2: Illustration of how violating the two explanatory desiderata will affect theory-development during later chiseling based on cognitive constraints. a) When the theory *assumes* the explanandum, instead of explaining it, it can be seen as hollow, which may be revealed during the chiseling process. b) When the theory is piecemeal—made up of several components to explain different aspects of the phenomenon, without forming a coherent whole—it may break apart upon further chiseling (green outline indicates target area to be chiseled off). Stone images taken from freepik.com

pute *tractable* functions for real-world input sizes (as opposed to toy scale). This means that for an explanation of a cognitive capacity to be plausible, it has to be computationally tractable (van Rooij, 2008; van Rooij, Blokpoel, Kwisthout, & Wareham, 2019). One can analyse whether a given theory is tractable by specifying it at Marr's computational level (as a function that maps from input to output) (Marr, 1982), and using mathematical proof techniques from computational complexity theory (van Rooij, 2008; van Rooij et al., 2019).

If the theory of interest turns out to not be tractable, similar techniques from *parameterized* complexity theory can be used to find out whether the input domain can be constrained in a way that *would* make it tractable (van Rooij, 2008; van Rooij et al., 2019). By input domain here we mean anything that is part of the input to the function that describes the *what* of the cognitive capacity. Note that this is a different notion of *input* than the sensory input to the neural or cognitive system when we process language (e.g., the auditory input of a speech stream, the visual input of a sign or gesture stream, etc.). Instead, the input domain in this context also includes any stored knowledge that is used in the explanation of how the cognitive system gets to a certain output (e.g., an interpretation of a novel expression), such as lexico-semantic knowledge, grammatical knowledge, world knowledge, etc. For an example of such a parameterized tractability analysis applied to a theory of intentional communication that involves inferring others' communicative goals, see van Rooij et al. (2011).

Neural plausibility Computational tractability is analysed at Marr's computational level of analysis (Marr, 1982), and thus only requires a computational-level model that describes the *what* of the cognitive capacity in question. That is, describing the nature of the input-output mapping being computed (the cognitive *function*). However, as Martin (2016) argues, any model of language computation must not only answer such *what* questions, but also *how* questions. That is, to provide a specification at the algorithmic level of analysis: describing the nature of the algorithmic process by which the cognitive function is being performed (the cognitive *process*). Similarly, Hagoort (2019) argues that the computational, algorithmic, and implementational level are interdependent, and that this should be taken into account when developing a mechanistic account of meaning-making in the mind (or in fact any cognitive function).

The set of possible algorithms is constrained by the computational-level explanation, but is also underdetermined by it (Blokpoel, 2018). That is, a given cognitive function (input-output mapping) can in principle be computed by different algorithms (van Rooij et al., 2019; van Rooij & Blokpoel, 2020; Blokpoel, 2018). However, as Martin (2016, 2020) demonstrates, algorithmic-level explanations can be constrained and informed by what we know about how the brain works: What type of computations can neural systems carry out? (e.g., summation and normalization.) (see also Martin, 2020; Kaushik & Martin, 2022). In addition to constraining possible theories to only those cognitive functions for which there exists an algorithm that can tractably compute it (see Section *Computational tractability*), one can further constrain the space of possible theories by putting additional constraints on the type of algorithm. Given a particular set of operators (e.g., summation and normalization) that are considered plausible for the brain to implement, one could make the commitment that the function needs to be computable by an algorithm that uses only these operators. In other words, one can make assumptions about the kind of architecture that cognition is implemented in, and make the commitment that the cognitive function a theory posits should be computable by an architecture of this type (van Rooij, 2008).

Challenges for explaining flexible linguistic inference

The sections above outlined the phenomenon to be explained, as well as the form that a good explanation should take, all together summarised in Figure 3. In the process of theory-development, tensions may arise between each of these properties and constraints. Figure 3 can thus be seen as describing a 'force field', within which tensions may arise both within and across levels. Below, we work out two of these tensions in a bit more detail: (i) explaining compositionality without assuming it (tension between a property of the phenomenon and an explanatory desideratum), and (ii) explaining the role of world knowledge tractably (tension between a property of the phenomenon and a cognitive constraint).

Explaining compositionality without assuming it

Explaining the compositional nature of language comprehension without assuming it raises questions for what type of linguistic knowledge can be considered part of the input domain (see Section *Computational tractability* for what we

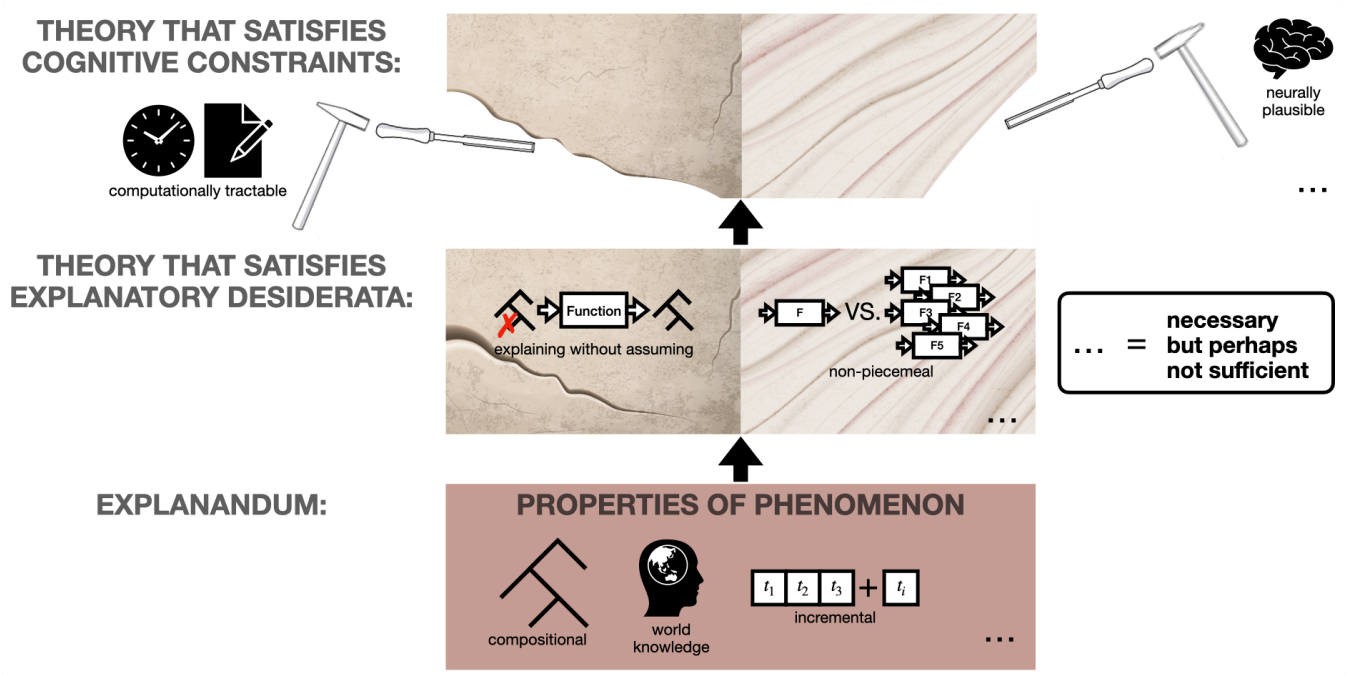

Figure 3: Illustration of how the explanandum and the constraints on the explanans relate to each other. The explanandum is characterised by three core properties. To explain this explanandum, a theory has to be able to capture these three properties (i.e., be *generatively sufficient*; Blokpoel, 2018). Evaluating the theory along the explanatory desiderata will help make sure that it really provides an explanation of the phenomenon of interest, and making sure the theory fits within the cognitive constraints will help make sure it can be plausibly realised by human cognition and the brain. Stone and tool images taken from freepik.com

mean by input domain). There is a tension between assuming that the relevant grammatical knowledge is in place (i.e., that we're explaining flexible linguistic inference in competent adult language users), and providing an explanation of the computational operations that are necessary to get from a linear input sequence to a hierarchical, compositional representation. The latter is what requires explanation, and this computational process *itself* cannot be assumed to be part of the input domain, else we are assuming without explaining.

Explaining the role of world knowledge tractably

Blokpoel et al. (2019) present a computational-level model of how novel candidate hypotheses may be generated through *deep analogical inference*: where structured representations of knowledge are (iteratively) related to each other through analogy, in a way that allows for augmentation of structured representations (through projection of parts from one representation to its analogous representation). Blokpoel et al. (2019) highlight several necessary properties for this model, one of which is *isotropy*: That all knowledge is potentially relevant in the inference process (see also Blokpoel, 2015, chapter 1; and Fodor, 1983, part IV). The explanandum we focus on in this paper is related to the explanandum of Blokpoel et al. (2019) in the sense that flexibly coming up with plausible interpretations of neologisms probably requires coming up with *novel* structured representations based on the combination of linguistic knowledge and world knowledge that is activated by the incoming expression. In fact, this type of flexible interpretation of novel communicative signals is exactly the example that Blokpoel et al. (2019) use to illustrate their explanandum. The question they ask is: How are candidate hypotheses generated in abductive inference? Where (a) plausible interpretation(s) of a novel communicative expression is an example of such candidate hypotheses. This raises issues for computational tractability, because if all world knowledge is potentially relevant, how can this component of the input domain be constrained? (See Section *Computational tractability* and Blokpoel et al., 2019, Section 4.2.)

Conclusion

Above, we worked out two examples of tensions that arise between different components of the force field that we identified in this paper. What we learn from these examples is that it is challenging even to satisfy one of the metatheoretical commitments that we put forward as important for explaining cognitive processes (i.e., the explanatory desiderata and cognitive constraints), while at the same time doing justice to each core property of the phenomenon (flexible linguistic inference) in its full capacity. Moreover, in the two examples above we limited ourselves to pairwise tensions between one property of the phenomenon and one metatheoretical commitment, but three-way or more-way tensions are also possible.

Other pairwise tensions that we did not have space to cover in this paper include: (i) explaining incremental compre-

hension in a non-piecemeal way (how to account for different levels of linguistic analysis?); (ii) explaining compositional comprehension in an neurally plausible way (compositionality requires symbolic processing—variable-value independence—while the brain excels at statistical and associative learning; Martin & Baggio, 2020); and (iii) compositional comprehension and the involvement of world knowledge (how are world knowledge and linguistic knowledge integrated?). We encourage theoreticians to work out three- or more-way tensions between the different properties and constraints we put forward in this paper. To conclude, explaining flexible linguistic inference while satisfying these properties and constraints (Figure 3) poses a major challenge. The theory-development needed to solve this challenge, requires a keen awareness of the force field we exposed in this paper.

Acknowledgments

The authors thank Olivia Guest and Laura van de Braak for insightful discussions on the explanandum, as well as metatheoretical commitments and the role they play in theory-development. We further thank Anna Mai, Cas Coopmans, Sophie Slaats, Jinbiao Yang, Xiaochen Zheng, Ashley Lewis, and Elena Mainetto for asking questions that helped us sharpen the ideas presented in this paper. We thank the people mentioned above and the other members of the Computational Cognitive Science (CCS) group at Donders Centre for Cognition, Radboud University, the Language and Computation in Neural Systems (LaCNS) group at the Max Planck Institute for Psycholinguistics and Donders Centre for Cognitive Neuroimaging, and the Big Question 5 team of the Language in Interaction Consortium for feedback on presentations of this work in lab meetings. We also thank three anonymous reviewers for their helpful feedback on an earlier version of this paper. MW and this research were supported by Big Question 5 (to Prof. dr. Roshan Cools & Dr. Andrea E. Martin) of the Language in Interaction Consortium, funded by NWO Gravitation Grant 024.001.006 to Prof. dr. Peter Hagoort. AEM was supported by a Max Planck Research Group and a Lise Meitner Research Group "Language and Computation in Neural Systems", and by NWO Vidi grant 016.Vidi.188.029.

References

Adolfi, F., van de Braak, L., & Woensdregt, M. (2023). *From empirical problem-solving to theoretical problem-finding perspectives on the cognitive sciences.* OSF.

Adolfi, F., Wareham, T., & van Rooij, I. (2023). A Computational Complexity Perspective on Segmentation as a Cognitive Subcomputation. *Topics in Cognitive Science*, *15*(2), 255–273.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198).

Blokpoel, M. (2015). *Understanding understanding: A computational-level perspective*. Unpublished doctoral dissertation.

Blokpoel, M. (2018). Sculpting Computational-Level Models. *Topics in Cognitive Science*, *10*(3), 641–648.

Blokpoel, M., Wareham, T., Haselager, P., Toni, I., & van Rooij, I. (2019). Deep Analogical Inference as the Origin of Hypotheses. *The Journal of Problem Solving*, *11*(1), 1–24.

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press.

Guest, O. (2024). What Makes a Good Theory, and How Do We Make a Theory Good? *Computational Brain & Behavior*.

Guest, O., & Martin, A. E. (2023). On Logical Inference over Brains, Behaviour, and Artificial Neural Networks. *Computational Brain & Behavior*, *6*(2), 213–227.

Hagoort, P. (2019). The meaning-making mechanism(s) behind the eyes and between the ears. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(20190301).

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*(5669), 438–441.

Hagoort, P., & van Berkum, J. (2007). Beyond the sentence given. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*(1481), 801–811.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354.

Kaushik, K. R., & Martin, A. E. (2022). *A mathematical neural process model of language comprehension, from syllable to sentence.* PsyArXiv.

Lahiri, A., & Marslen-Wilson, W. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition*, *38*(3), 245–294.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. MIT Press.

Martin, A. E. (2016). Language Processing as Cue Integration: Grounding the Psychology of Language in Perception and Neurophysiology. *Frontiers in Psychology*, *7*(120).

Martin, A. E. (2020). A Compositional Neural Architecture for Language. *Journal of Cognitive Neuroscience*, *32*(8), 1407–1427.

Martin, A. E., & Baggio, G. (2020). Modelling meaning composition from formalism to mechanism. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(20190298).

McQueen, J. M. (2007). Eight questions about spoken word recognition. In M. G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics.* Oxford University Press.

Partee, B. H. (1995). Lexical semantics and compositionality.

In *Language: An invitation to cognitive science, Vol. 1, 2nd ed* (pp. 311–360). MIT Press.

Pylkkänen, L. (2020). Neural basis of basic composition: what we have learned from the red–boat studies and their extensions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(20190299).

Sedivy, J. C. (2007). Implicature During Real Time Conversation: A View from Language Processing Research. *Philosophy Compass*, *2*(3), 475–496.

Slonimska, A., Özyürek, A., & Capirci, O. (2020). The role of iconicity and simultaneity for efficient communication: The case of Italian Sign Language (LIS). *Cognition*, *200*(104246).

Szabó, Z. G. (2004). Compositionality. In *Stanford Encyclopedia of Philosophy.*

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science*, *268*(5217), 1632–1634.

van Rooij, I. (2008). The Tractable Cognition Thesis. *Cognitive Science*, *32*(6), 939–984.

van Rooij, I. (2022). Psychological models and their distractors. *Nature Reviews Psychology*, *1*(3), 127–128.

van Rooij, I., & Blokpoel, M. (2020). Formalizing Verbal Theories. *Social Psychology*, *51*(5), 285–298.

van Rooij, I., Blokpoel, M., Kwisthout, J., & Wareham, T. (Eds.). (2019). *Cognition and intractability: a guide to classical and parameterized complexity analysis*. Cambridge University Press.

van Rooij, I., Guest, O., Adolfi, F. G., de Haan, R., Kolokolova, A., & Rich, P. (2023). *Reclaiming AI as a theoretical tool for cognitive science.* OSF.

van Rooij, I., Kwisthout, J., Blokpoel, M., Szymanik, J., Wareham, T., & Toni, I. (2011). Intentional Communication: Computationally Easy or Difficult? *Frontiers in Human Neuroscience*, *5*.

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Processes*, *56*(3), 229–255.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185.