This is the penultimate draft of a paper forthcoming in *Linguistics and Philosophy*. Please cite the final version when it becomes available

Eye-tracking evidence for the causal-historical theory of reference

Filippo Domaneschi – University of Genoa, filippo.domaneschi@unige.it
Nicolò D'Agruma – Vita-Salute San Raffaele University, nico.dagruma@gmail.com
Massimiliano Vignolo – University of Genoa, massimiliano.vignolo@unige.it
Camilo R. Ronderos – University of Oslo, camilo.rodriguez.ronderos@gmail.com

Abstract

In this paper, we present an experiment that shows conflicting findings from truth-value judgments and eye-tracking data for testing reference assignment of proper names. We argue that if eye-tracking is a more reliable method than truth-value judgment tasks, then our eye-tracking data provide stronger empirical support for Kripke's causal-historical theory of reference for proper names. We also argue that eye-tracking and truth-value judgments cannot both be reliable techniques for resolving the debate. If they were, they should yield convergent results. Instead, we find that the truth-value judgment data align with the descriptivist prediction, while the eye-tracking results conform to the Kripkean pattern.

Keywords:
Causal Theory of Reference
Descriptivist Theory of Reference
Experimental semantics
Eye-tracking
Proper names

1. Introduction

The intense experimental debate on the methodology of theories of reference of proper names, namely on the criteria for determining the correctness of one theory over another, started in 2004 with the publication of "Semantics, Cross-Cultural Style" by Machery, Mallon, Nichols, and Stich. The debate has led scholars to focus on truth-value judgments (TVJs) in Gödel/Schmidt-like cases. In this paper, we present an eye-tracking experiment that shows that there is a clash between the method of truth-value judgments in Gödel/Schmidt-like cases—which is one of the most widely adopted but also highly controversial methods in experimental philosophy on theories of reference of proper names—and the method of eye-tracking for testing reference assignment. We argue that, if eye-tracking is a reliable method for testing reference assignment, the method of truth-value judgments in Gödel/Schmidt-like cases is not a reliable method for testing theories of reference. Moreover, under the assumption that eye-tracking is a reliable method for testing reference assignment, the findings of our experiment provide evidence for the causal-historical theory of reference.

The plan of the paper is as follows. In section 2, we summarize the debate that led scholars to focus on the method of truth-value judgments in Gödel/Schmidt-like cases. In section 3, we summarize the ambiguity objection that has been raised against the reliability of truth-value judgments. In section 4, we explain why the question regarding the reliability of truth-value judgments is critical for a theory of reference. In section 5, we discuss eye-tracking and how it has been used in psycholinguistics for studying reference assignment and language comprehension. In sections 6 and 7, we present the findings of our eye-tracking experiment that are in conflict with truth-value judgments. In section 8, we draw our conclusion about the methodology of experimental philosophy on theories of reference of proper names.

2. Reference and truth-value judgments

Machery, Mallon, Nichols, and Stich (2004) (henceforth MMNS) launched an impressive attack on Kripke's refutation of classical descriptivism and the methodology of philosophers in theory of reference. MMNS assumed (i) that Kripke provided his main argument against classical descriptivism with the Gödel/Schmidt case and (ii) that the Gödel/Schmidt case is an application of the method of cases.[1] With respect to theories of reference, the method of cases states that if a theory of reference predicts that a proper name N refers to an object O, the theory is confirmed if laypeople have the intuition that N refers to O and disconfirmed if they

[1] See MMNS (2004, p. B3): «There is widespread agreement among philosophers on the methodology for developing an adequate theory of reference. The project is to construct theories of reference that are consistent with our intuitions about the correct application of terms in fictional (and nonfictional) situations».

have the intuition that N does not refer to O. The reliability of the method of cases presupposes the uniformity of referential intuitions. MMNS argued against the uniformity of referential intuitions and claimed that philosophical theories of reference are methodologically baseless.

MMNS argued that people's referential intuitions vary depending on demographic factors: Westerners appeared to have referential intuitions that are more in line with the causal-historical theory, while East Asians seemed to have referential intuitions that are more in line with classical descriptivism.[2] MMNS tested 31 undergraduates from Rutgers University (USA) and 40 from Hong Kong University (China). Participants were laypeople, namely speakers who lacked any specialized philosophical expertise. MMNS presented two scenarios modeled after Kripke's Gödel/Schmidt case to each experimental group.[3] One case used the proper name "Gödel", the other used the proper name "Tsu Ch'ung Chih". This is the scenario with the proper name "Gödel":

> Suppose that John has learned in college that Gödel is the man who proved an important mathematical theorem, called the incompleteness of arithmetic. John is quite good at mathematics and he can give an accurate statement of the incompleteness theorem, which he attributes to Gödel as the discoverer. But this is the only thing that he has heard about Gödel. Now suppose that Gödel was not the author of this theorem. A man called "Schmidt", whose body was found in Vienna under mysterious circumstances many years ago, actually did the work in question. His friend Gödel somehow got hold of the manuscript and claimed credit for the work, which was thereafter attributed to Gödel. Thus, he has been known as the man who proved the incompleteness of arithmetic. Most people who have heard the name "Gödel" are like John; the claim that Gödel discovered the incompleteness theorem is the only thing they have ever heard about Gödel.
>
> When John uses the name "Gödel", is he talking about:
> (a) the person who really discovered the incompleteness of arithmetic? or
> (b) the person who got hold of the manuscript and claimed credit for the work?
> (MMNS, 2004, p. B6)

[2] MMNS make their prediction based on a body of research in cognitive psychology, such as Norenzayan et al. (2002), Nisbett (2003), and Nisbett et al. (2001). For a criticism against the basis of MMNS's prediction, see Martí (2012), Devitt (2012b), and Ostertag (2013).

[3] MMNS's experiment also included two vignettes testing the Jonah case (Kripke, 1980, p. 67). However, the experimental literature has mainly focused on the Gödel scenario because MMNS's Jonah cases failed to elicit results in line with the cross-cultural hypothesis.

Classical descriptivism predicts (a), while the causal-historical theory predicts (b). Table 1 reports the results.[4]

	Percentage
Gödel/Schmidt-like case with Western name	
Westerners	58
East Asians	29
Gödel/Schmidt-like case with East Asian name	
Westerners	55
East Asians	32

Table 1 Percentages of intuitions in line with the causal-historical theory in MMNS (2004)

MMNS detected a significant cross-cultural variation. Furthermore, the findings also indicated a noteworthy intra-cultural variation. Importantly, several experimental works replicated MMNS's effects, by emphasizing an intra and cross-cultural variation on proper names' reference (for an overview, see Machery 2021, 2024; for a meta-analysis, see van Dongen et al., 2021; for some remarks on the meta-analysis, see Machery, 2024, p. 190).

Devitt (2011, 2012a, 2012b, 2015a) and Martí (2009, 2012, 2014, 2020) criticized MMNS's study. They claimed that the reality that researchers need to observe to test theories of reference is *linguistic usage*, that is how people use proper names. They argued that experimental studies have implications for theories on reference if those studies show that laypeople use proper names as predicted by classical descriptivism or by the causal-historical theory. MMNS's study failed to provide any evidence in that regard, since it did not show how participants use proper names. Rather, it collected participants' meta-linguistic intuitions on proper names, which places the experimenters one step removed from the reality that scholars should test, namely, actual linguistic usage.

Moreover, Martí—but not Devitt—argues that those meta-linguistic intuitions are theoretically biased, as participants likely expressed them by reflecting on how reference is determined, relying on their theoretical inclinations toward the causal-historical theory or classical descriptivism.[5] It is not surprising that many participants expressed a descriptivist intuition. The idea that a name designates the entity satisfying the associated description has its credibility. After all, influential philosophers such as Russell and arguably Frege proposed

[4] Machery and Stich (2012) reported the percentage formulation of the results, while MMNS (2004) presented their findings by using a less straightforward format.

[5] It must be noted that Martí's objection that questions about reference invite participants to theorize about reference itself is confined to the logical space as there is no empirical evidence in support of it.

and defended classical descriptivism. As said, Devitt too contends that scholars should test theories of reference against usage; however, he (2011, p. 432, n. 14) disagrees with Martí regarding the supposed theoretical bias inherent in referential intuitions. While some participants may respond to MMNS's prompt by theorizing about reference, Devitt stresses that the prompt asks about an instance of reference and not about theories of reference. Therefore, *per se*, that is not a theoretical prompt. According to Devitt, participants' intuitions are not the right evidence not because they are biased, but because they are one step removed from the reality that theories of reference are about, namely linguistic usage, and there is no guarantee that laypeople's intuitions reliably reflect that reality.[6]

Machery, Olivola, and de Blanc (2009) (henceforth MOD) replied to Martí's (and Devitt's) criticism that scholars should test linguistic usage and conducted an experiment with participants from India, France, and Mongolia. MOD presented a case like the one employed in MMNS (2004) with the name "Tsu Ch'ung Chih". In each group, one subset of participants performed the same prompt as in MMNS (2004). The other subset of participants answered the following question:

> Having read the above story and accepting that it is true, when Ivy says, "Tsu Ch'ung Chih was a great astronomer", do you think that her claim is: (A) true or (B) false?
> (MOD, 2009, p. 690)

This is a truth-value judgment test, in which a participant has to evaluate whether what the speaker says by uttering a sentence is true or false. According to MOD, the *rationale* of the task is this: the speaker's inclination to judge the utterance of the sentence as true indicates her inclination to assert the uttered sentence, so her disposition to use that sentence, hence her linguistic usage. The sentence "Tsu Ch'ung Chih was a great astronomer" is true only if "Tsu Ch'ung Chih" refers to the discoverer of the solstice times. Therefore, the answer "True" would show that the participant is willing to use the name "Tsu Ch'ung Chih" as predicted by classical descriptivism. Instead, the sentence is false if "Tsu Ch'ung Chih" refers to the thief

[6] It is important to stress that not all authors engaged in the debate agree with Martí's and Devitt's criticism. In fact, beyond the experimental reply formulated by Machery et al. (2009), which we are about to detail in the main text, Machery (2011, 2012, 2014) argues that Martí's and Devitt's hypothesis lacks empirical support and contradicts the usual methodology in linguistics. Our work is potentially interesting for both the scholars who consider referential intuitions reliable and the scholars who do not. The former can view the forthcoming presentation of our eye-tracking experiment as introducing a new and potentially valuable methodology to add to those already used in the literature; the latter can see our work as offering a potentially better-equipped technique compared to referential intuitions. We will assume this dialectical caveat throughout the paper, including when discussing the methodologies in Domaneschi and Vignolo (2020) and Li (2021), as well as when presenting the eye-tracking data themselves.

of the astronomical discovery. Hence, the answer "False" would show that the participant is willing to use the name as predicted by the causal-historical theory. Table 2 reports MOD's results, with additional data from a group of participants from the USA (presented by Machery and Stich, 2012).

Group	**TVJ**	**RI**
India	67	58
France	56	48
Mongolia	66	56
USA	64	67

Table 2 Percentages of truth-value judgments (TVJ) and referential intuitions in line with the causal-historical theory in MOD (2009) and Machery and Stich (2012)

The TVJs results corroborate the intra-cultural variation, as each group presents a significant percentage of participants providing the answer predicted by classical descriptivism. While the TVJ percentages predicted by the causal-historical theory are higher than the corresponding referential-intuition percentages (except for the USA sample), the differences are small and do not reach statistical significance—that is, the two series of values are congruent. MOD inferred a cross-cultural variation across Westerners and Easterners through inductive reasoning. Given the above congruence between TVJs and referential intuitions, the cross-cultural difference that MMNS (2004) reported would have also emerged if Westerners and Easterners had been asked to express TVJs (i.e., to perform a test on language usage). As a consequence, MOD claimed that Martí's objection to MMNS's (2004) experiment is not warranted.[7]

We do not agree with the reasons that MOD provided for the claim that TVJs test linguistic usage. In fact, the circumstance that a participant judges the utterance of the sentence "Gödel was a great mathematician" ("Tsu Ch'ung Chih was a great astronomer") by John (Ivy) as true does not entail that she is inclined to assert that sentence. Unlike John, the participant reads a vignette, thanks to which she acquires some information that is inaccessible to John. As a consequence, the participant gets familiar with a new description that may be associated with "Gödel", such as "The mathematician who stole the proof of the incompleteness theorem to take the credit for it". Therefore, according to descriptivism, the description guiding the

[7] The inductive hypothesis that MOD formulated finds more direct corroboration with Li et al. (2018) and Li's (2021) Experiment 1, who presented some TVJ Gödel-style cases to American and Chinese participants, in English and Mandarin, and detected a cross-cultural variation.

participant's use of the name may differ from the description guiding John's use of the name. In other words, also according to the descriptivist theory, the participant may use the proper name to refer to the thief of the theorem, and not to the discoverer.[8]

The reasons why TVJs do test linguistic usage are different. As several authors have pointed out (Cohnitz and Haukioja, 2013, 2015, 2021; Maynes, 2015; Devitt and Porot, 2018; Vignolo and Domaneschi, 2018; Domaneschi and Vignolo, 2020), TVJs shed light on the participant's *understanding* of John's utterance of the sentence "Gödel was a great mathematician" (or Ivy's utterance of "Tsu Ch'ung Chih was a great astronomer"). Usage of proper names unfolds not only in the production of sentences containing them, but also in the comprehension of them when they occur in sentences uttered by other speakers. To echo Cohnitz's words: «presumably, production and interpretation are just two sides of the same coin» (2015, p. 96). If a participant judges John's utterance of the sentence "Gödel is a great mathematician" as true, her TVJ reveals that she understood John's utterance to express a proposition about Schmidt. That is evidence about linguistic usage (comprehension) and not about referential intuitions. As we explain in section 4, if the method of TVJs is reliable, the TVJ "True" is evidence from linguistic usage against Kripke's refutation of classical descriptivism. This circumstance would give TVJs a prominent methodological role in testing theories of reference. Before explaining why that is so, it is helpful to discuss an objection that has been raised against the reliability of TVJs and the reasons why that objection is not conclusive.

3. The ambiguity objection against truth-value judgments

As Domaneschi and Vignolo (2020) emphasized, TVJ tests effectively capture linguistic comprehension only if the experiment controls for an ambiguity in the interpretation of the truth predicate. Domaneschi and Vignolo argued that the predicate "True" can be understood to mean *true according to the speaker's epistemic perspective*, i.e. *true according to what the speaker believes*, or *true according to the narrator's epistemic perspective*, i.e. *true according to the facts reported in the vignette*. Only the narrator's epistemic perspective elicits TVJs that are apt to test the two theories of reference.[9]

[8] This criticism aligns with the one that Devitt and Porot (2018) term the "New Meaning Objection". While Vignolo and Domaneschi (2018, pp. 8–9, n. 7) and Domaneschi and Vignolo (2020, p. 445, n. 8) endorse that objection (albeit not in a fully developed way), Devitt and Porot (2018) contend that the New Meaning Objection is misconceived. Discussing Devitt and Porot's arguments for their claim falls outside the scope of this paper.

[9] As Domaneschi and Vignolo (2020) explain, the epistemic ambiguity that they identify in the context of TVJs is a development of a similar ambiguity that Sytsma and Livengood (2011) identify in the context of referential intuitions (see also Sytsma et al., 2015). Also Li (2021) maintains that participants can perform the TVJ tests according to two different epistemic perspectives, but, unlike Domaneschi and Vignolo, Li claims that both the

The point that distinguishes classical descriptivism from the causal-historical theory is that John's utterance of the sentence "Gödel was a great mathematician" (or "Tsu Ch'ung Chih was a great astronomer") is factually true according to classical descriptivism and factually false according to the causal-historical theory. John's utterance, however, is true from the perspective of what John believes according to both theories. Therefore, only if the truth predicate is understood as meaning *true according to the facts (reported in the vignette)*, the TVJ "True" is evidence that participants understand that John's utterance is about Schmidt, who is the referent predicted by classical descriptivism.

Three observations are in order at this point. First, Domaneschi and Vignolo (2020) do not explicitly equate the "narrator's perspective" with the factual truth, although they confirm that the equivalence underpinned that paper.

Second, the label "narrator's perspective" that Domaneschi and Vignolo adopt may be somewhat misleading. While the participant adopting John's perspective judges whether John's utterance aligns with what John believes, the participant judging the factual truth value does not evaluate whether John's utterance aligns with the narrator's or anyone else's beliefs. Rather, the participant judges whether John's utterance reflects the facts themselves. Domaneschi and Vignolo are right that a participant can judge the factual truth value of John's utterance only if she relies upon the entire account that the narrator provides. Yet, one matter is the information that the participant uses to formulate her TVJ; while another matter is the nature of the TVJ itself: if a participant goes beyond what John believes, then her TVJ concerns the facts, that is, the objective reality.

Third and relatedly, one may question the notion of "ambiguity" as applied to TVJs. To use an example from Devitt and Porot (2018, p. 1555), if one asks whether (it is true that) it rained at Trump's inauguration, one's question is not about anyone's opinion (i.e., perspective) about the rain, but rather about a fact—whether it indeed rained or not. Therefore, participants would not oscillate between two equally valid interpretations of an ambiguous predicate. Rather, some participants understand the truth predicate literally ("factually true"), while other participants misunderstand it non-literally ("true from John's

---

speaker's perspective and the narrator's are irrelevant to testing theories of reference. However, we find Li's (2021) conceptualization of the narrator's perspective problematic. While space constraints prevent us from fully elaborating on our reasons here, the issue with Li's *external condition*, which experimentally operationalizes the narrator's perspective, lies in the final prompt, which replaces John (or Ivy) with a new omniscient character, Kermit. That approach eliminates the crucial element of disagreement between descriptivist and causal-historical theories, that is, the truth value of a statement *as uttered by a speaker unaware of certain crucial events*. For instance, John's unawareness of the theft in the Gödel case leads him to associate "Gödel" only with "The discoverer of the theorem". For a more comprehensive criticism of Li (2021) and also of the developments in her subsequent work (Li, 2023a), where she ultimately claims that both the speaker's and the narrator's perspectives, as operationalized in her 2021, provide relevant data for testing theories of reference, see D'Agruma (2023).

perspective").[10] However, Reuter and Brun (2022) and Reuter (2024) have provided experimental data suggesting that the truth predicate can indeed be ambiguous, along lines similar to the ones that Domaneschi and Vignolo suggest. We will remain neutral on the issue. That said, in the forthcoming sections, we will adhere to Domaneschi and Vignolo's term "ambiguity" for the sake of simplicity, as our arguments do not hinge on whether one considers the phenomenon to be one of genuine ambiguity. The interpretation of "truth" relevant to theories of reference and experimental testing is the "factual" one, irrespective of whether that is one of several meanings literally associated with the truth predicate or the only literal one. No one denies that John is uttering a sentence that he believes to be true; the disagreement lies in whether John is making a factually true statement, which is the case only according to the descriptivist theory.

Domaneschi and Vignolo (2020) conducted an experiment in order to disambiguate the interpretation of the truth predicate.[11] They tested a group of Italian speakers with an Italian translation of the simplified version of the Super Dog Race that Li et al. (2018 p. 109) employed in their experiment:

> Long ago, there was a race called the Super Dog Race. Max, Pickles and Blaze participated in the race. Max crossed the finish line first, winning the race, but he got too excited and ran all the way to the North Pole. Pickles crossed the finish line second. He stopped and watched Max run away. The announcer of the race mistakenly thought that Pickles won the race. He told every newspaper in the world that Pickles won. He also told them that another dog, Blaze, ran very fast despite his short legs. Since then, everyone learned that Pickles won the race. They don't know anything else about Pickles.
>
> Tom and Emily learned at school that Pickles won the Super Dog Race. This is the only thing they know about the dog race and Pickles. They don't know anything about Max.
>
> (Domaneschi and Vignolo, 2020, pp. 448–9)

[10] Devitt (2023, pp. 1151–4) raises similar observations while discussing the various readings of the truth predicate that Li (2023a, pp. 65–8, 92–127) proposes. For Li's reply to Devitt, see Li (2023b).

[11] Domaneschi and Vignolo (2020) conducted three studies. However, the study that we present—Experiment 1—is the main one, while the other two—Experiment 2 and 3—play more of a supporting role.

Participants were asked to perform a truth-value judgment:

> That night, their dad asked: Do you know who won the Super Dog Race?
> Emily said: "Pickles was the dog that won the Super Dog Race".
> Do you think that her claim is
> (a) true or
> (b) false?
> (Domaneschi and Vignolo, 2020, p. 449)

The participants who chose "True"—the purportedly descriptivist answer—were then asked a follow-up question. The aim was to ascertain the epistemic perspective that they adopted when providing the TVJ.

> You think Emily's claim is true because:
> (c) Pickles did not win the Super Dog Race, but Emily believes that Pickles won the race because her teacher told it to her.
> (d) What Emily believes is a true description of Max and she uses the name "Pickles" to talk about the dog that really (unbeknown to Tom and Emily's teacher) won the Super Dog Race.
> (Domaneschi and Vignolo, 2020, p. 449)

If the participants who judge Emily's utterance as true do so because they understand Emily's use of the name with the referent predicted by classical descriptivism, then they should choose (d), for the choice of (c) means that they interpret the truth predicate with the meaning *true according to what Emily believes*. Domaneschi and Vignolo observed that 42% of participants answered that Emily's utterance is true. However, 75% of them answered (c) to the follow-up question. Thus, the vast majority of those participants provided a TVJ that cannot be taken to be evidence for classical descriptivism. Domaneschi and Vignolo concluded that the ambiguity of the truth predicate impacted on MOD's experiment and testing TVJs in Gödel/Schmidt-like cases is not a reliable method for testing theories of reference.

Domaneschi and Vignolo's objection is not conclusive. The problem is that their experiment rests upon a follow-up question, which is arguably susceptible to the criticism that Martí and Devitt raised against the tests of referential intuitions. For that question asks participants to provide a justification for their TVJ, potentially distancing experimenters from

the raw linguistic-usage data against which theories of reference should be tested. More specifically, participants in Domaneschi and Vignolo's (2020) experiment may answer the follow-up question by choosing the explanation that they find more intuitive and easier to understand. The follow-up presents two different options to justify the TVJ "True": (c) centers on what Emily believes, while (d) relies upon the supposed referential connection between the name "Pickles" and the real winner of the race. In Domaneschi and Vignolo (2020) participants may choose the explanation focusing on Emily's perspective because it appears easier to understand, even if it does not accurately reflect the genuine reason behind their TVJ. In other words, a descriptivist participant may provide the TVJ "True" on genuinely descriptivist grounds and nonetheless she may select the follow-up answer (c) because she finds that explanation more intuitive and easier to process. The participant may select option (c) over the more convoluted (d), which involves difficult expressions like "unbeknown to Tom and Emily's teacher", "true description of", and "[Emily] uses the name".

Machery (2021) expressed another concern about Domaneschi and Vignolo's (2020) experiment. His criticism is that asking participants to justify their answer may give them a cue that their answer is wrong and push them to give a causal-historical justification. Machery's objection provides an additional reason to question Domaneschi and Vignolo's methodology.[12]

Li (2021) conducted a study—Experiment 1—structurally similar to Domaneschi and Vignolo's Experiment 1 (2020).[13] Li presented some Gödel/Schmidt-like scenarios to Western (American) and Eastern (Chinese) participants, who performed a TVJ task on sentences like "Gödel is the discoverer of the incompleteness theorem". Li's results corroborate MMNS's cross-cultural variation across the two groups: across the four vignettes that she uses, Westerners and Easterners chose the TVJ predicted by the causal-historical theory, namely "False", 90% and 32% of the time respectively[14]. After expressing their TVJs,

[12] See Machery (2021, p 545): «unfortunately, Experiment 1 suffers from severe demanding characteristic: asking participants Question 2 [i.e., the follow-up question] invites them to rethink their answer to Question 1 [i.e., the TVJ prompt]. So, even if participants had descriptivist leanings when answering Question 1, they would be pushed to give a causal-historical answer when answering Question 2».

[13] Li's Experiment 1 involves three conditions in a between-subjects design: the *original condition*, the *external condition*, and the *internal condition*. These conditions are structurally comparable to Domaneschi and Vignolo's Experiments 1, 2, and 3, respectively (albeit with some relevant differences: as mentioned, we take Li's *external condition* not to operationalize the narrator's perspective properly). Just as we focus on Domaneschi and Vignolo's Experiment 1 because it is the prominent study, we similarly consider Li's *original condition* as arguably the main one and thus we present only that condition. Li (2021) also includes Experiment 2, in which she administers the three conditions in a within-subjects design to Chinese participants only, replicating the results obtained from Chinese participants in Experiment 1.

[14] It is worth stressing that Li (2021)—like Li et al. (2018)—did not ask participants whether the sentence is true or false, but whether the speaker "is right" or not to utter it, under the assumption that a speaker is right if and

participants were asked to justify it by answering an open follow-up question. Li found that the participants who expressed the TVJ "True"—namely the purportedly descriptivist answer—justified that choice by appealing to the speaker's epistemic perspective—what the speaker believes to be true—analogously to what happened in Domaneschi and Vignolo (2020). Li's choice of using an open follow-up question may offer a partial solution to the problem that we spotted in Domaneschi and Vignolo (2020). While in Domaneschi and Vignolo (2020) the two follow-up options differ in terms of the complexity of their phrasing, Li's (2021) participants were free to formulate their justification.[15] However, Li's methodology remains problematic because it still relies upon a follow-up question, which moves the experiment one step away from the linguistic-usage data, thereby reviving Martí's and Devitt's criticism. It is conceptually possible[16] that a participant gives her justification by appealing to the explanation that she finds simpler, not the explanation that accurately describes her actual understanding of the utterance.[17]

4. Truth-value judgments: a challenge to Kripke

In this section, we argue that the reliability of TVJs is critical for theory of reference. If TVJs are reliable, i.e. if they are not affected by the ambiguity of the truth predicate discussed in the previous section, then MOD's findings present a threat to Kripke's refutation of classical descriptivism. In particular, we point out that such a threat cannot be ignored even by scholars like Deutsch (2009, 2010, 2015) who reject the view that intuitions have a dominant evidential role in the methodology of philosophical theories of reference.

Before explaining why TVJs (if reliable) are so crucial, two remarks are in order. First, we want to avoid any misunderstanding that could arise from the philosophical implications of the notion of intuition. According to many scholars, intuitions are inclinations to make judgments (Devitt 2006, p. 103; Machery 2017, p. 38). Thus, in what follows, when we speak of *referential intuitions,* we mean *dispositions to make referential judgments*. Second, we are

---

only if her utterance is true and she is not if and only if her utterance is false. One may doubt the appropriateness of such an equivalence; however, discussing this aspect falls outside the scope of this paper, especially since the TVJ task included in our eye-tracking experiments uses the classical, arguably less controversial, true/false question.

[15] Naturally, the advantage deriving from the open format of the follow-up question comes with the cost of obtaining less easily analyzable data, since coding participants' explanations as originating from the adoption of the speaker's perspective or not may pose challenges.

[16] This is a mere speculation with no empirical support and the objection is confined to the logical space.

[17] One may argue that, irrespective of the ambiguity of the TVJ "True", the strikingly high incidence of American participants selecting the TVJ "False" (90%) overrides any possible concern regarding the validity of the causal-historical theory for this demographic. That conclusion would be hasty. Other TVJ results involving American or more broadly Western participants (Machery and Stich, 2012; Li et al., 2018; Vignolo and Domaneschi, 2018; Domaneschi and Vignolo, 2020) show some degree of oscillation, with all the sample percentages for the TVJ "False" remaining distant from the 90% threshold.

not siding with Deutsch's view. We just want to stress that if TVJs are reliable, they challenge Kripke's refutation of classical descriptivism also for scholars who embrace Deutsch's reconstruction of Kripke's Gödel/Schmidt case and take referential intuitions to be irrelevant to the confirmation or disconfirmation of philosophical theories of reference. Such scholars can question the debate that led to focus on TVJs, but cannot avoid facing the findings of TVJs.

Kripke's Gödel/Schmidt case refutes classical descriptivism by using a counterexample. It falsifies the descriptivist thesis that the reference of a proper name is determined by uniquely identifying information that speakers associate with it. If the proper name "Gödel" refers to Gödel in the Gödel/Schmidt case, then classical descriptivism is false. There is broad consensus that the justification for the claim that the proper name "Gödel" refers to Gödel in the Gödel/Schmidt case comes from the fact that philosophers have the referential intuition (i.e., disposition to make the referential judgment) that "Gödel" refers to Gödel in the Gödel/Schmidt case. Call it the *Gödel Intuition*.[18] The fact that philosophers have the *Gödel Intuition* is evidence that its content is true, namely that it is true that "Gödel" refers to Gödel in the Gödel/Schmidt case.

As said in the previous sections, experimentalists reacted critically to this methodology and aimed to show that a significant percentage of Western speakers and a large majority of Eastern speakers do not have the *Gödel Intuition*. These results (if reliable) leave Kripke's refutation of classical descriptivism in trouble. Actually, they leave all philosophical theories of reference in trouble, if the methodology of those theories essentially relies on the evidential role of referential intuitions collected with the method of cases.

Deutsch argues that the reconstruction of Kripke's Gödel/Schmidt case as an argument that relies on the evidential role of referential intuitions collected with the method of cases is mistaken.[19] According to Deutsch, Kripke provided an argument for justifying the conclusion that the proper name "Gödel" refers to Gödel in the Gödel/Schmidt case that does not rest upon the fact that philosophers have the *Gödel Intuition*.

Deutsch observes that in *Naming and Necessity* Kripke introduced the Gödel/Schmidt case after the presentation of real-life cases in which speakers are not able to associate a proper name with any uniquely identifying description and, nonetheless, they are able to refer to the bearer of the name. For example, many people know of Cicero only that he was a famous Roman orator and of Feynman that he was a famous physicist. Still, they refer to Cicero when

[18] Here we follow Deutsch (2015, p. 8) who speaks of the *Gödel Judgment*.
[19] Cappelen (2012) agrees with Deutsch's metaphilosophical view and extends it to philosophy in general. For a reply to Deutsch and Cappelen, see Devitt (2015b).

using the proper name “Cicero” and to Feynman when using the proper name “Feynman”. These real-life cases falsify the descriptivist thesis that being able to associate a proper name with a uniquely identifying description is a necessary condition for referring to the bearer of the name. Kripke employed the Gödel/Schmidt case to address the descriptivist thesis that being able to associate a proper name with a uniquely identifying description is a sufficient condition for referring to the satisfier of the description.

Deutsch also observes that, immediately after the presentation of the Gödel/Schmidt case[20], Kripke presented three real-life cases of speakers who associate proper names with uniquely identifying descriptions that are not satisfied by the bearers of the proper names. Despite that, speakers refer to the bearers of the proper names, not to the satisfiers of the descriptions. Many speakers believe that Peano was the author of the axioms of the theory of natural numbers, but he was not. Many speakers believe that Einstein was the inventor of the atomic bomb, but he was not. And many speakers believe that Columbus was the first European to land in the Western hemisphere, but he was not. Nonetheless, those speakers refer to Peano with “Peano”, to Einstein with “Einstein”, and to Columbus with “Columbus”.

In Deutsch’s reconstruction (Deutsch 2015, pp. 11–2), the comparison with real-life cases provides justification for the claim that the proper name “Gödel” refers to Gödel in the hypothetical Gödel/Schmidt case. The proper name “Gödel” refers to Gödel in the hypothetical Gödel/Schmidt case by analogy with real-life cases in which speakers refer to the bearers of the proper names independently of the descriptions they associate with them. Whatever the referential relations in the real-life cases happen to be, the same referential relation obtains in the hypothetical case too. The fact that philosophers have the *Gödel Intuition* does not matter to the justification of its content, i.e. that “Gödel” refers to Gödel in the Gödel/Schmidt case.

According to Deutsch, then, the Gödel/Schmidt case is an argument by analogy with the *Ignorance and Error Argument*, as named by Devitt and Sterelny (1999). The *Ignorance and Error Argument* is based on the observation of real-life cases in which speakers use proper names to refer to their bearers even if they are not able to associate them with uniquely identifying descriptions or if the descriptions they associate them with are not true of the bearers of the proper names. The *Ignorance and Error Argument* rests upon premises such as, for example, that speakers who associate the description “The discoverer of the axioms of the theory of natural numbers” with the proper name “Peano” refer to Peano, not to Dedekind.

[20] Deutsch (2015, p. 13) maintains that Kripke presents the Gödel/Schmidt case for addressing the point that even in those cases in which we already know that the descriptions associated with proper names identify their bearers, it is not because the names are associated with true descriptions that they refer to their bearers.

As Devitt (2020, pp. 433–7) points out, scholars who assign an evidential role to referential intuitions reinstate that methodological role at this stage for justifying the premises of the *Ignorance and Error Argument*.[21] The Gödel/Schmidt case does not appeal to referential intuitions but to analogy with the real-life cases of the *Ignorance and Error Argument*, but referential intuitions are nonetheless needed for giving justification to the premises of the *Ignorance and Error Argument*.[22]

Deutsch, who denies the evidential role of referential intuitions in Kripke's refutation of classical descriptivism, rejects this reconstruction of the *Ignorance and Error Argument*. According to him, the *Ignorance and Error Argument* starts with the factual observation that many speakers do not associate proper names with descriptions that uniquely identify their bearers. Thus, if classical descriptivism were true, the uses of proper names such as "Cicero", "Feynman", "Columbus", "Peano", and "Einstein" by ignorant or mistaken speakers would fail to refer to their bearers. For this reason, classical descriptivism gives us a relation of reference that implies that many proper names, as used by ignorant or mistaken speakers, do not refer to their bearers. According to Deutsch, Kripke thought that this implication of classical descriptivism is falsified by evidence from linguistic usage. For example, according to classical descriptivism, the referent of "Peano", for speakers who associate "The discoverer of the axioms of natural number theory" with the name, is whoever is uniquely identified by that description. That implies that such speakers speak the truth if they utter, "Peano was the discoverer of the axioms of natural number theory." If classical descriptivism were true, it would be impossible to account for utterances of the sentence "Peano was the discoverer of the axioms of natural number theory" by ignorant or mistaken speakers as expressing false content about Peano. According to Deutsch, in the *Ignorance and Error Argument*, Kripke took it from granted that evidence from linguistic usage shows that ignorant or mistaken speakers who utter sentences like "Peano was the discoverer of the axioms of natural number theory" are treated as competent and are understood as speaking falsely by their audience.[23]

---

[21] Devitt (2020, p. 435-438) argues that the referential intuitions that count for confirming or disconfirming theories of reference are those of experts, i.e. philosophers, and there is empirical evidence, part of it coming from experiments on linguistic usage, that laypeople's referential intuitions are not reliable.

[22] It must be noted that Machery et al. (2013) stress that laypeople's referential intuitions on humdrum cases might be in line with the rejection of classical descriptivism that is Kripke's goal in *Naming and Necessity*. Yet, they argue that referential intuitions on hypothetical cases like the Gödel/Schmidt one are necessary for advancing from the rejection of a given theory of reference to the individuation of the correct theory of reference. They point out that their theoretical goal is not to criticize Kripke's rejection of classical descriptivism but to raise doubts on the philosophical enterprise of constructing any theory of reference by showing that it is a methodologically flawed enterprise.

[23] It is assumed that Kripke observed this kind of empirical data from linguistic usage as a member of his linguistic community. As Martí (2012, 2014) suggests, philosophers of language are both competent speakers and objective observers capable of maintaining the kind of distance required to reflect on the linguistic practice within their community. This is not to say that philosophers do not need to study empirically how speakers use

This is evidence from linguistic usage that their utterances of the sentence "Peano was the discoverer of the axioms of natural number theory" do not have the content predicted by classical descriptivism.

In Deutsch's reconstruction (2015, p. 14), this empirical claim about how ignorant and mistaken speakers are understood is crucial for justifying the premises of the *Ignorance and Error Argument*. In his view, the empirical claim (i.e., that speakers are understood as referring to the bearers of the proper names even if they do not associate uniquely identifying descriptions with them) is crucial for justifying through an inference to the best explanation the conclusion that their uses of the names are acts of referring to the bearers of the names (i.e., that their uses of "Cicero", "Feynman", "Columbus", "Einstein", and "Peano" refers to Cicero, Feynman, Columbus, Einstein, and Peano, respectively). Now, it should be clear why TVJs are so important even for scholars who endorse this interpretation of the *Ignorance and Error Argument*. If MOD's findings on TVJs are reliable, they show that a significant amount of laypeople take John, the hypothetical character who plays the role of a mistaken speaker in the vignette, to tell the truth when he utters a sentence like "Gödel was a great logician". These results seem to falsify to a significant extent the empirical claim that serves to justify the premises of the *Ignorance and Error Argument* in Deutsch's reconstruction.

If TVJs are reliable, they cast a doubt on Kripke's refutation of classical descriptivism. More generally, they cast doubt on the methodology of philosophical theories of reference. Such a treat stands also for those scholars who reject the evidential role of referential intuitions collected with the method of cases.

5. Eye-tracking

We explained why the question of the reliability of TVJs is critical for philosophical theories of reference. We also pointed to empirical work doubting the reliability of TVJs (Domaneschi and Vignolo, 2020; Li, 2021). However, in these cases, participants were asked to provide justifications that problematically rely on meta-linguistic judgments concerning names and their referential properties. In order to truly assess the reliability of TVJs, we would need to compare their results with those from a task that does not rely on meta-linguistic judgments.

To overcome this problem, Cohnitz and Haukioja (2015) and Cohnitz (2015) suggested using the eye-tracking, visual-world paradigm to probe language comprehension, a method

proper names and language in general. It means that philosophers can access empirical data on linguistic usage without making experiments, but simply as competent speakers of their linguistic community. However, empirical data collected simply as members of linguistic communities might not be sufficient or might be wrong. In fact, if TVJs are reliable, Kripke's alleged observations of linguistic usage are false to a significant measure. It is false to significant measure that ignorant or mistaken speakers are understood as referring to the bearers of the names they use. And this shows the need for empirical data collected with experiments.

that has successfully been used in psycholinguistics throughout the last three decades (see Tanenhaus et al., 1995; Huettig et al., 2011). In a typical eye-tracking, visual-world study, participants see images that are potentially related to the incoming linguistic input that they hear via speakers or headphones (e.g., by depicting referents to specific words). Their eye movements toward said images are then recorded. In this way, researchers can explore (among other things) the referential connections that a participant establishes between the words she listens to and the referents she sees pictured in front of her (Tanenhaus and Spivey-Knowlton, 1996).

Previous research using this method has in fact shown that listeners rapidly and automatically assign referents to words during incremental language comprehension (e.g., Altmann and Kamide, 1999, 2007; Tanenhaus et al., 1995). Further research has also shown that participants can rapidly anticipate upcoming objects based on hearing verbs with specific selectional restrictions (Altmann and Kamide, 1999). Moreover, several studies have shown that contextual information influences the assignment of reference for a verb's argument (e.g., Kamide et al., 2003; Chambers and San Juan, 2008; Ronderos et al., 2023). The crucial advantage of this method is that participants do not need to formulate any explicit judgments, providing experimenters with the type of linguistic-usage data that Martí and Devitt view as critical.

Moreover, eye movements are less dependent on theoretical reflection of the kind that Martí stresses, as they are natural, spontaneous, and, for the most part, involuntary. In contrast, TVJs only measure linguistic comprehension *after* the comprehension process has occurred. The participant *first* hears and understands an utterance and *then* evaluates it for truth or falsity. During the gap between the two stages, theoretical reflections may intervene and influence the TVJ. Instead, eye-tracking sheds light on a participant's linguistic comprehension *while* the understanding process unfolds, that is, while the participant is listening to an utterance. Eye-tracking drastically reduces the time lag between comprehension and data collection, preventing participants from consciously considering the two possible explanations—description-based and causal-based—and expressing their intuitive inclinations toward either. Admittedly, we cannot rule out that some form of theoretical bias could implicitly affect the eye-tracking data collection if participants consider the two possible explanations in a non-overt manner. However, ruling out conscious engagement with theories is a significant step toward minimizing the potential theoretical influence that Martí highlights.

As Cohnitz and Haukioja's (2015) state, eye-tracking allows researchers to test the reference of proper names «without having to ask them [i.e., participants] anything at all», that is, without the need to ask them to make overt judgments about the critical utterance of the sentence they are presented with (p. 638). Inspired by Cohnitz's (2015) and Cohnitz and Haukioja's (2015) insights, we therefore conducted an eye-tracking study. In the following sections, we present our study and its results.

6. Experiment

In our experiment, we used the eye-tracking, visual-world paradigm to better understand how people spontaneously assign reference in Gödel/Schmidt cases. We recorded participants' eye movements while they heard sentences containing a proper name while simultaneously inspecting pictures of the name's potential referents. The working hypothesis in eye-tracking is that the picture participants look at while listening to a word is the picture of the referent they assign to it. Assuming that the way participants understand words is part of their linguistic usage as much as their linguistic production is, eye-tracking results provide linguistic usage data for testing theories of reference.

We also collected TVJs to understand the degree to which these align with the eye-tracking record. Assuming that eye-tracking is a reliable test of language comprehension, if participants look at the picture of the referent predicted by the causal-historical theory (e.g. Gödel), they should answer that what the speaker says by uttering a sentence like "Gödel is the discoverer of the incompleteness theorem" is false, since they understand the speaker to refer to Gödel with the name "Gödel" (and they are told that Gödel was a thief). If participants look at the picture of the referent predicted by classical descriptivism, Schmidt, they should answer that the utterance is true, since they understand the speaker to refer to Schmidt with the name "Gödel" (and they are told that Schmidt is the author of the incompleteness theorem). If instead participants look at the picture of the referent predicted by the causal-historical theory and answer that the utterance is true, then, under the assumption that eye-tracking is a reliable method, this would be evidence that they do not interpret the truth predicate as meaning *true according to the facts* (reported in the vignette). A plausible hypothesis in line with previous studies is therefore that participants interpret the true predicate as meaning *true from the speaker's epistemic perspective*, i.e. *true with respect to what the speaker believes*. The methodological bottom line is the following. Assuming that eye-tracking is a reliable way to test how participants spontaneously assign reference, the extent to which other tests (such as TVJs) are congruent with the eye-tracking record equals

the extent to which said tests provide reliable data regarding the spontaneous assignment of reference.

Method and Design

50 Italian speakers [31f; MA=25.50; SD=3.52] were presented with 10 vignettes (randomized and in Italian) modeled after the Gödel/Schmidt case. Specifically, the vignettes were all stories in which a speaker, like everyone else in her linguistic community, associates a name with a description identifying an individual different from the bearer of the name as used in the vignette. While we kept the vignettes' structure constant, we varied the content across the stories: for instance, some stories involved a theft (like in the classical Gödel scenario), while others presented a case of involuntary misattribution. Some stories involved intellectual achievements (such as the discovery of a theorem), while others were related to sporting accomplishments, and so on. No filler or control items were included to avoid excessive cognitive demands on participants. All the vignettes' potential referents—which we call "Target" and "Competitor"—were human beings. Figure 1 presents an example. Participants read the first part of the vignette (*Step 1*), introducing the causal-historical referent (Target) and a picture of it (*Step 2*). Participants then read the second part of the vignette (*Step 3*), introducing the descriptivist referent (Competitor) and a picture of it (*Step 4*). Participants read the final part of the vignette (*Step 5*). After a two-second pause (*Step 6*), four pictures (Target, Competitor, and two distractors) appeared on the screen (*Step* 7), randomly distributed across the four quadrants of the screen: to clearly differentiate between the two potential referents (Target and Competitor) and the two distractors, the two referents were always of the same gender (i.e., male or female), while the distractors were always of the opposite gender. Participants listened to a critical sentence uttered by the speaker (the character of the vignette, e.g. John) while they viewed the four images on the screen. At this point, we recorded their eye movements toward the Target and Competitor images using a Tobii Pro X3-120 eye-tracker. The experiment had a single factorial predictor, 'position', with two levels: FIRST and LAST.

In the FIRST condition, the critical sentence was as follows: "Gödel is the discoverer of the incompleteness theorem". In the LAST condition, the sentence was as follows: "The discoverer of the incompleteness theorem is Gödel". In a within-subject design, each participant was presented with target sentences from the FIRST condition in five vignettes and from the LAST condition in the other five. The order of the vignettes was randomized, while the allocation to the FIRST or LAST condition was semi-randomized across the 10 vignettes, with each participant receiving one of two possible combinations. The rationale is

to verify whether, while hearing the proper names, participants looked at the Target or at the Competitor images, even in the LAST condition where the definite description had a prenominal position and the property of being *The discoverer of the incompleteness theorem* is more salient.

After listening to the sentence, participants were asked to provide their truth-value judgments by answering the following question: "…do you believe that his/her [*the character of the vignette*] claim is true or false?" (*Step 8*). The materials, raw data, and analysis scripts are available on the project's OSF repository. Materials are given in both the original Italian version and an English translation (the translation is literal, following the lexical and syntactical choices of the Italian text as closely as possible).[24]

[24] At this link: https://osf.io/kgnte/?view_only=e94bb99ae08847f7b0a3631535be20f6.

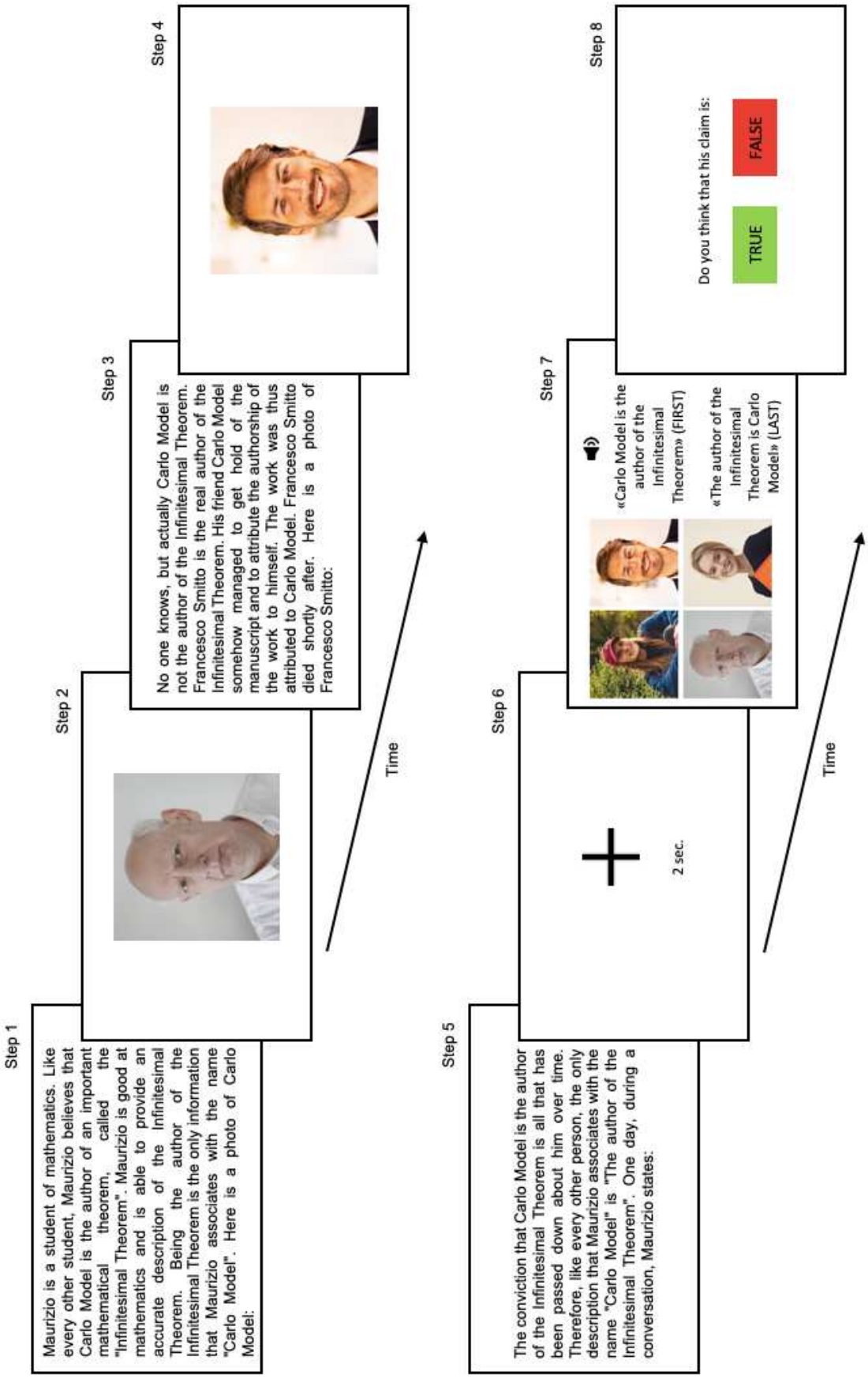

Fig. 3 Example of vignette (translated into English) with its steps

Analysis

To analyze the truth-value judgments, we calculated the percentages of "True" and "False" responses across participants in the two conditions (FIRST and LAST). We then fitted a mixed-effects, logistic regression model to the data, including random intercepts and slopes by items and by participants.

To analyze the eye-tracking record, we calculated the proportions of looks to Target, Competitor, and distractors while participants listened to the sentences. We examined only eye movements of those participants who answered "True" in the truth-value judgment task. We focused only on those participants because, as explained in sections 3 and 4, the response "True" is the kind of response that is supposed to provide evidence against Kripke's refutation of classical descriptivism and has been taken to be affected by the ambiguity of the truth predicate.[25] Proportions of looks were calculated for every 20 milliseconds time bin across two pre-determined time-windows: (i) the "critical" region, time-locked from the onset until the offset of the proper name, and (ii) the "description" region, time-locked from onset to offset of the rest of the sentence. We then fitted mixed-effects linear regression models to each time-window, including random intercepts and slopes by items and by participants. The dependent measure used was log-gaze probabilities of looks to the Target image divided by looks to the Competitor image (Arai et al., 2007). This measure is commonly used in several studies employing the visual-world paradigm (for references, see Barr, 2008; Huettig et al., 2011; Magnuson, 2019; Ito and Knoeferle, 2023) and allows us to quantify whether there is a significant preference to look at the Target or Competitor image in each of the two time-windows. The models had the factor "Position" as a predictor, which was treatment contrast coded. We fitted each model twice, changing the condition coded as the baseline (FIRST or LAST). This was critical to test whether the intercept of the model was significantly different form zero in each case, which amounts to testing whether there was a preference for looking at the Target image (positive numbers) or the Competitor image (negative numbers) in each of the conditions.[26]

[25] The ambiguity applies only to the TVJ "True" because John (or Ivy or whoever else) takes his utterance to be true, and this circumstance may incline participants to formulate that TVJ. As a consequence, one may wonder whether the collected TVJs "True" concern the factual reality or the speaker's perspective. In contrast, the TVJ "False" does not present a similar ambiguity. Since John does not take his utterance to be false (e.g., he does not intend to deceive), the TVJ "False" can stem only from the consideration of the factual truth.

[26] As a reviewer correctly notes, the fixations calculated in each time-bin are typically not independent, since they are correlated with the time-bins that come before and after (given that the duration of a fixation is typically longer than 20 ms and will therefore be represented across multiple time bins). This autocorrelation of observations is a problem for studies that are concerned with the temporal emergence of an effect, since the residuals of one time-bin could predict the residuals of the next adjacent one (thus violating the assumption of the linear model, see Ito and Knoeferle, 2023). However, in the present study, we are not concerned with the temporal emergence of the effect and we rather focus on overall effects that are visible across the entire region.

Results

The regression model for the truth-value judgment task failed to find a statistically significant difference between the truth-value judgments across the two conditions ($p = 0.19$, z-value = 1.29) (Fig. 2). In both conditions, participants answered that the sentence is true about 85% of the time—i.e., 86% in FIRST and 84% in LAST, while participants answered "False" 14% and 16% of the time in FIRST and LAST respectively.

The results of the eye-tracking analysis are shown in Figure 3 and 4. The model for the critical region corresponding to hearing the proper name (Fig. 3) showed a significant difference between conditions ($p < 0.001$, t-value = 3.573, ß= -0.46133). Re-fitting the model with different baseline conditions showed that the FIRST condition (where the name precedes the description) was positive and significantly different from zero ($p < 0.001$, t-value = 6, ß = 0.35), while the LAST condition (where the description precedes the name) was negative and not significantly different from zero ($p = 0.3$, t-value = 0.9, ß = -0.1). The model for the region corresponding to hearing the description (Fig. 4) also showed a significant difference between conditions ($p < 0.001$, t-value = 7.3, ß = 0.98). Re-fitting the model with different baseline conditions showed that the FIRST condition was positive and significantly different from zero ($p < 0.001$, t-value = 7.8, ß = 0.6), while the LAST condition was negative and significantly different from zero ($p < 0.001$, t-value = 3.8, ß = -0.3).

Concerning the distractor images, Figure 5 and 6 report the overall proportion of looks throughout both regions for all four images while participants hear the proper name. As can be seen, in both cases, the proportion of looks to the distractor images is close to zero in the majority of the time, with the exception of the early parts of the sentence, in which the proportion of looks to all of the images hovers around chance (i.e., 25%). This signals that (i), there were no baseline differences between conditions, and all effects are caused by hearing the sentence unfold, and (ii) distractors were not considered to be viable referents to the proper name.

Though a more fine-grained analysis that provided detail on the temporal dynamics of the effects could be informative, it is outside of the scope of the current investigation, where we are interested in showing that the overall results of a TVJ test and the eye-tracking record diverge in terms of testing theories of reference.

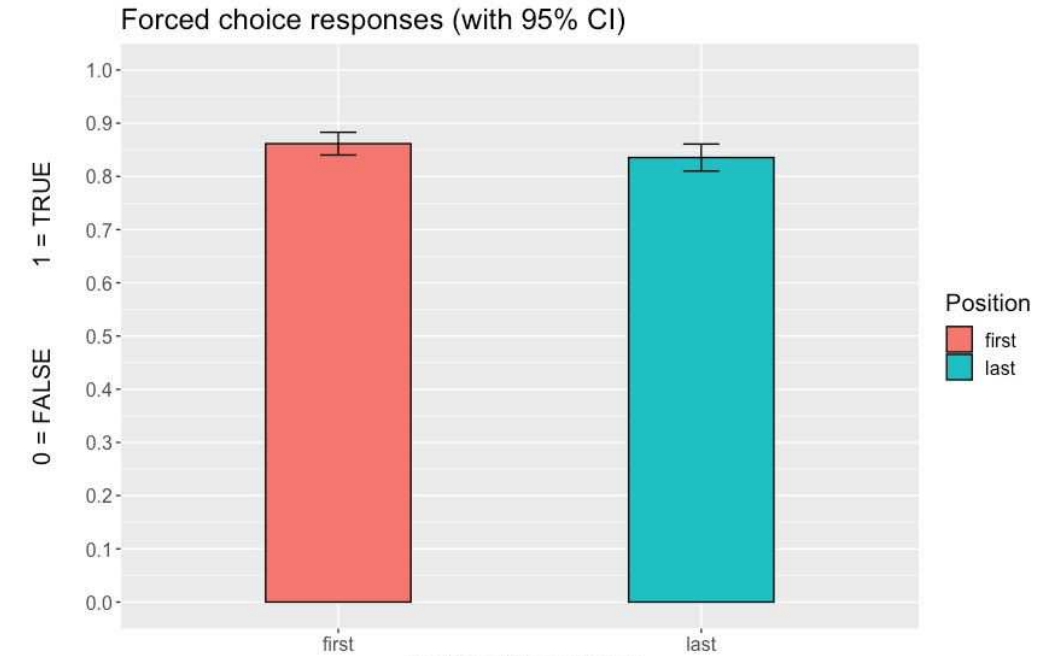

Fig. 2 Proportions of "True" and "False" answers

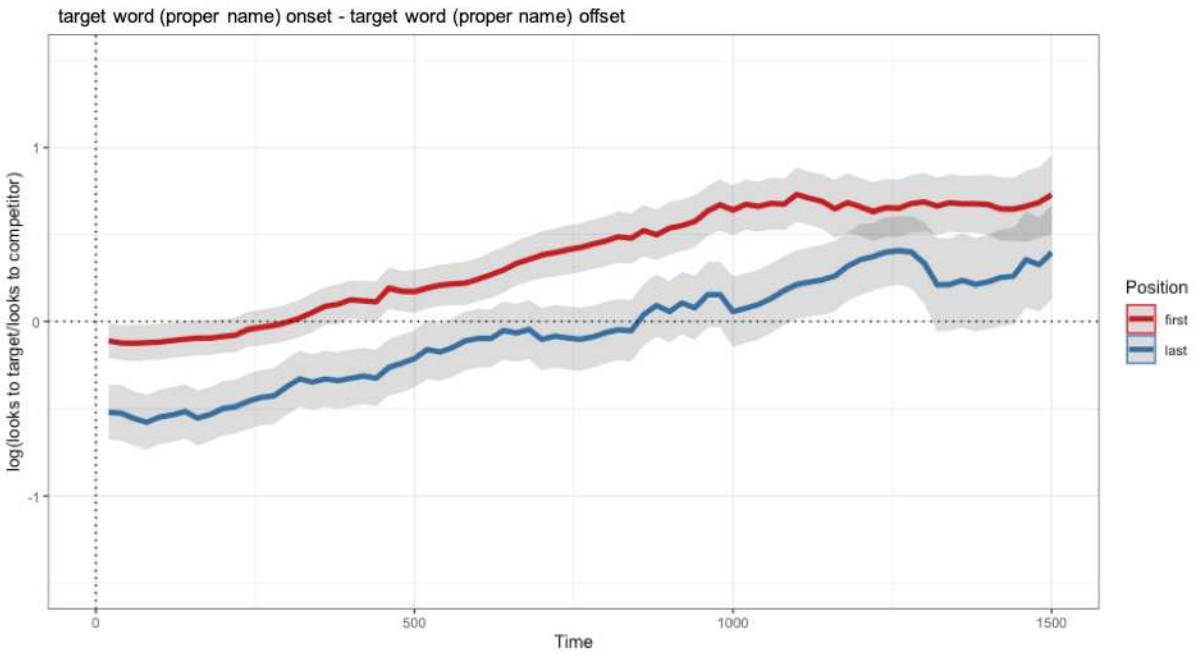

Fig. 3 Log-gaze probabilities of looks to the Target image divided by looks to the Competitor image, time-locked from onset to offset of the proper name time-window. Positive numbers represent looks toward the Target image, while negative numbers represent looks toward the Competitor image. Grey ribbons show confidence intervals.

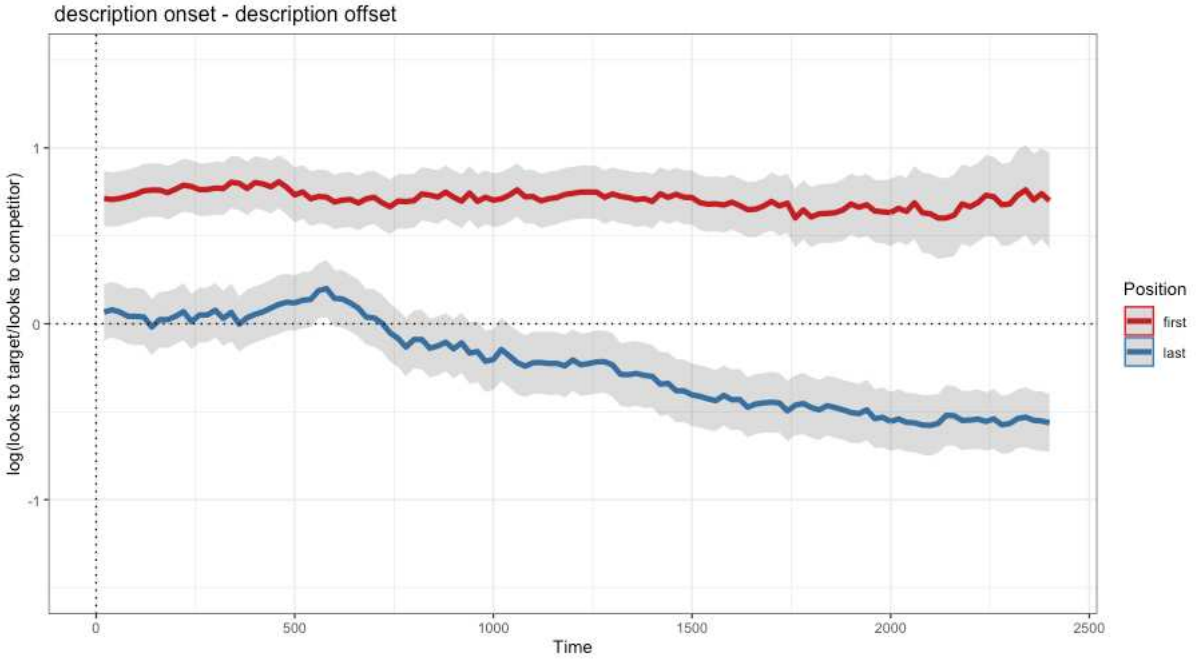

Fig. 4 Log-gaze probabilities of looks to the Target image divided by looks to the Competitor image, time-locked from onset to offset of the description time-window. Positive numbers represent looks toward the Target Image, while negative numbers represent looks toward the Competitor image. Grey ribbons show confidence intervals.

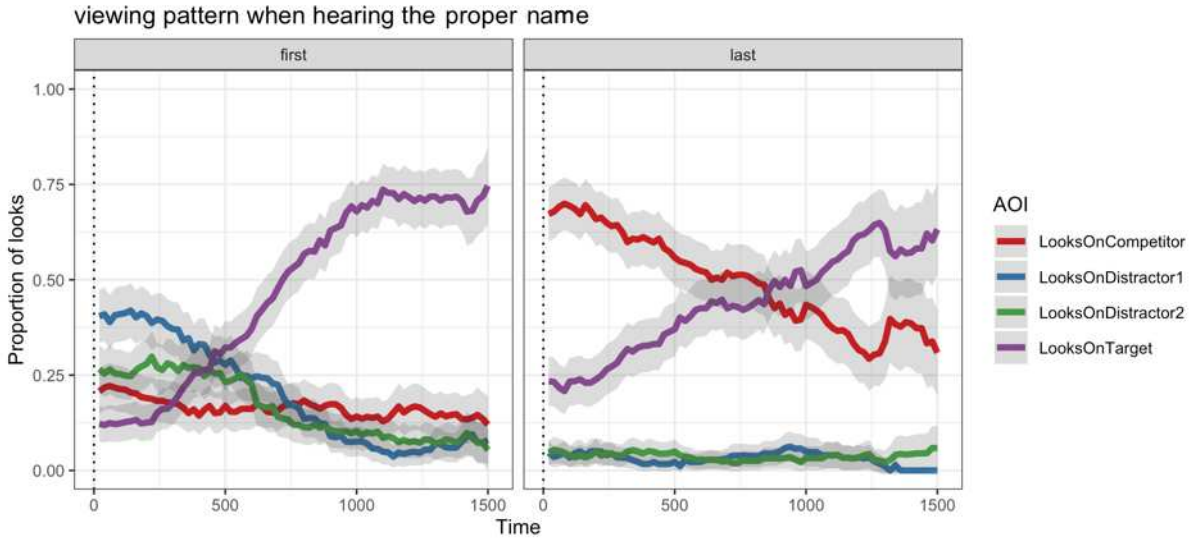

Fig. 5 Log-gaze probabilities of looks to the Target image, the Competitor image, and the two distractors, when participants hear the proper name.

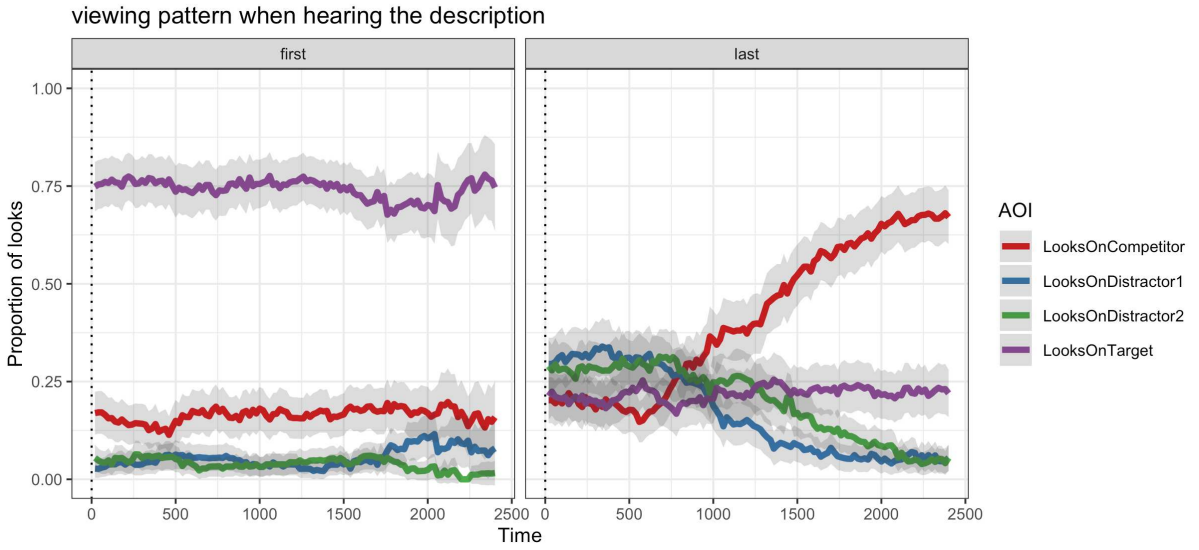

Fig. 6 Log-gaze probabilities of looks to the Target image, the Competitor image, and the two distractors, when participants hear the definite description.

7. Discussion

The eye-tracking data provided support to the claim that participants understood that the speaker used the proper name for referring to the referent predicted by the causal-historical theory. In the FIRST condition, participants heard the speaker uttering, for example, the sentence "Carlo Model is the author of the Infinitesimal Theorem". When participants listened to the first part of the sentence, namely the proper name "Carlo Model", they looked at the Target image, as is evident from the red line in Figure 3, which is significantly positioned above zero. Interestingly, participants continued looking at the Target image even when they listened to the second part of the sentence (the definite description "The author of the Infinitesimal Theorem") even if they knew that the Competitor image satisfied the description, as is evident from the red line in Figure 4, which shows how there were significantly more looks to the Target image than to the Competitor image in the FIRST condition. This eye-tracking pattern is evidence that participants understood that the speaker used (i) the proper name for referring to the Target image, the referent predicted by the causal-historical theory, and (ii) the definite description for the property of being *The author of the Infinitesimal Theorem* as predicate of the Target image.

In the LAST condition, participants heard the speaker uttering the sentence "The author of the Infinitesimal Theorem is Carlo Model". When participants listened to the first part of the sentence, namely the definite description "The author of the Infinitesimal Theorem", they looked at the Competitor image, as is evident from the blue line in Figure 4, which is significantly below zero. This is an expected result because participants knew that

Competitor is the satisfier of the description and the description takes the grammatical subject position of the sentence. However, the eye-tracking pattern becomes more interesting when we examine the data regarding the second part of the sentence, namely the proper name "Carlo Model". Participants started exhibiting a different trend: their gaze shifted from the Competitor image to the Target image, as illustrated in Figure 3. This happens to such an extent that the overall pattern represented by the blue line in Figure 3 is not significantly different from zero. In other words, upon hearing the proper name in the second part of the sentence, the participants shifted from looking at the Competitor image toward looking at the Target image, namely the individual that the causal-historical theory predicts. It could be speculated that the length of the utterance of the proper name did not afford participants enough time to consistently redirect their gaze toward the Target image. However, the important result is that the shift from the Competitor to the Target image shows that also in the LAST condition there was a higher tendency to look at the Target image while listening to the proper name.[27]

The results of the TVJs point in the opposite direction. In our experiment, the percentage of answers that the speaker's claim is true was around 85%. This suggests a discrepancy between data from TVJs and data from eye-tracking. The eye-tracking results show that participants understood the speaker to be using the proper name to refer to the referent predicted by the causal-historical theory, but data from TVJs show that participants tended to judge the speaker's claim to be true, as if they had understood the speaker to be referring to the referent predicted by classical descriptivism, not to the referent predicted by the causal-historical theory.[28]

---

[27] Notice that when participants hear the definite description in grammatical subject position (LAST), they look at the satisfier of the description (Competitor). Since Competitor is the semantic content of the description, this is evidence that, in the LAST condition, participants' understanding of the utterance of the target sentence is guided by the semantic interpretation of the uttered sentence. Since there is no structural difference between the LAST condition and the FIRST condition, we are justified to infer that in the FIRST condition too the participants' understanding of the utterance of the target sentence is guided by the semantic interpretation of the uttered sentence.

[28] A possible concern about the formulation of the vignettes is that Target and the corresponding picture always appear first, while Competitor and the corresponding picture appear only later. One might object that, in Step 7, participants tend to look at the Target image because that represents the first individual to whom they are introduced. Consequently, the vignettes could be biased in favor of the causal-historical theory. An indirect, partial response to this concern comes from Beebe and Undercoffer's (2016) Study 2a, where the authors modify the Gödel and Tsu Ch'ung Chih cases by altering the order of presentation of the characters compared to the seminal MMNS (2004) study. While MMNS first introduce the causal-historical referent and then the descriptivist one, Beebe and Undercoffer reverse the order. Beebe and Undercoffer replicate the variation that MMNS indicated. Therefore, in the context of their referential-intuition experiment, the order of presentation of the characters does not seem to exert a significant impact. Naturally, whether the order of presentation could influence our specific design, which involves pictures and the collection of eye-tracking data, remains an open question, which we leave for further research.

One may object that our material, by following the Gödel/Schmidt-like cases, is rather complex in terms of story and length, which may have affected the eye-tracking record. However, previous psycholinguistic work using the visual-world paradigm has been successfully used to study a wide range of phenomena, some of which require lengthier vignettes to ensure that the participant has generated specific expectations when faced with the critical stimulus (as was the case in the current experiment). In fact, the very first study to ever use the visual-world paradigm involved lengthy and multi-sentence stories that participants followed while freely inspecting images (Cooper, 1974). A prominent example much closer to the current investigation is the work by Ferguson and Breheny (2011). They showed participants two-sentence vignettes that set up expectations regarding an agent's hidden desires (e.g., "Tom doesn't want anyone to know his favorite color is pink. Last week, Tom bought a new car and he surprisingly chose a pink car"). By doing so, they found that, immediately upon hearing the critical adverb ("surprisingly"), participants looked at the picture that contradicted the agent's stated desires. This shows that participants do not only keep track of complex situations, but they can also rapidly use this information as soon as it becomes relevant to predict upcoming linguistic input. Similar findings (in terms of participants rapidly integrating complex, multi-sentence contextual information with incoming linguistic information) have been found, for example, in studies on metaphor processing (Ronderos et al., 2023) and on the processing of negative expressive adjectives (Ronderos and Domaneschi, 2023).

8. A conclusion on theories of reference

The challenge by MMNS (20004) on the methodology of philosophical theories of reference started a debate that led scholars to question the reliability of TVJs in Gödel/Schmidt-like cases. TVJs are supposed to provide data on how people use language. We explained that TVJs do provide data on linguistic usage, though not on language production but on language comprehension. More precisely, TVJs in Gödel/Schmidt-like cases provide data on how people understand proper names when used by other speakers. We explained that the findings of previous studies on TVJs raise a challenge to Kripke's refutation of classical descriptivism that cannot be ignored, not even by those philosophers who reject the view that intuitions have a core evidential role in the methodology of theories of reference. Those philosophers deny that in *Naming and Necessity* Kripke employed the Gödel/Schmidt case for polling referential intuitions. They claim that the Gödel/Schmidt case is an argument by analogy with the real-life cases that Kripke employed to mount the *Ignorance and Error Argument*. The crucial premise in the *Ignorance and Error Argument* is that many ordinary speakers refer to

the bearers of the proper names they use even though they are not able to associate a uniquely-identifying description of their bearers with them. The philosophers who deny recourse to intuitions justify this premise by an inference to the best explanation. Uses of proper names by ignorant or mistaken speakers are best explained as acts of referring to the bearers of the proper names. The empirical assumption in this explanation is that speakers who are not able to associate correct descriptions with proper names are nonetheless understood by other speakers to be referring to the bearers of the proper names. The findings of previous studies on TVJs seem to falsify this empirical assumption to a significant extent. The consequence would be a challenge to Kripke's refutation of classical descriptivism.

One way to uphold Kripke's refutation of classical descriptivism is to question the reliability of TVJs in Gödel/Schmidt-like cases. This work has been previously done in two studies (Domaneschi and Vignolo 2020, Li 2021) that show how TVJ tests are affected by the ambiguity of the truth predicate. Most of the participants who answered that what the speaker says is true interpreted the truth predicate as meaning *true according to what the speaker believes*. Under this interpretation of the truth predicate, the circumstance that a participant answers that the speaker's utterance is true is not evidence that the participant understands the speaker to be referring to the individual satisfying the speaker's description and not to the individual predicted by the causal-historical theory. In short, it is not evidence against the empirical assumption grounding the *Ignorance and Error Argument*.

We conducted an eye-tracking experiment to test how people understand proper names when used by other speakers in Gödel/Schmidt-like cases. The results of our experiment show that participants looked at the picture of the referent predicted by the causal-historical theory while listening to speakers using proper names in Gödel/Schmidt-like cases. Assuming that eye-tracking is a reliable method for testing language comprehension, these results suggest that participants understand the speakers to be referring to the individuals predicted by the causal-historical theory. Under this interpretation, they know that what the speakers say is false, because they learn it by reading the vignette. In spite of that, the results on TVJs show that participants tend to respond that what the speakers say is true.

Given our findings, it is possible to draw one of two conclusions that differ in their strength. The first conclusion is conditional. If one assumes that eye-tracking is a reliable technique in the context of Gödel-Schmidt scenarios, our findings support the case against the reliability of TVJs. This poses a significant challenge to a considerable body of work in

experimental philosophy concerning theories of reference of proper names.[29] Moreover, under the assumption that eye-tracking is a reliable technique in the context of Gödel-Schmidt scenarios, the results of our experiment count as evidence in favor of Kripke's refutation of classical descriptivism. They show that mistaken speakers are understood as referring to the bearers of the proper names predicted by the causal-historical theory of reference. This is evidence from linguistic usage (comprehension) in favor of Kripke's refutation of classical descriptivism.

Alternatively, given the discrepancy between the TVJ results and the eye-tracking ones, one could conclude that if one method is reliable regarding participants' spontaneous reference assignment preferences, the other one cannot be. If both techniques were dependable and informative in the same way, they should yield consistent results. However, that is not the case, as the TVJ results align with the descriptivist predictions, while the eye-tracking ones conform to the causal-historical predictions.

In the context of this paper, we opt for the latter, more cautious conclusion and leave the proof of the reliability of eye-tracking in the context of Gödel-Schmidt scenarios for future studies. That said, we recall that two previous studies (Domaneschi and Vignolo, 2020; Li, 2021) raise questions regarding TVJ—although, as discussed in section 3, there are some methodological controversies about those studies. This reinforces our confidence that future studies with eye-tracking will support the stronger conclusion against the reliability of TVJs and in favor of Kripke's refutation of classical descriptivism.

As a final note, it is important to highlight a key difference between our study and the existing literature. The percentage of answers "True" by Western participants has been previously reported to be around 40% (e.g., MOD, 2009; Machery and Stich, 2012; Vignolo and Domaneschi, 2018; Li et al., 2018; Domaneschi and Vignolo, 2020). In Li's (2021) Experiment 1 the percentage is as low as 10% and in Li's (2023a) Study 3 it oscillates between 12% and 33%—in both cases with American samples. In our experiment, this percentage is around 85%. Therefore, our results and Li's are somehow opposite: while her Western participants displayed a strong tendency to express the causal-historical answer, our results showcase the opposite trend. Moreover, it is noteworthy that, in the existing literature,

[29] The thesis that the eye-tracking technique is reliable encompasses not only the claim that eye-tracking is, in principle, an appropriate methodology to study names' reference, but also the claim that our specific implementation of the eye-tracking technique is flawless and experimentally sound. That includes aspects such as the wording of the vignettes, the choices of pictures and the correctness of the statistical analyses. Therefore, the reliability of the eye-tracking technique is a matter that also involves these more specific, implementation-related aspects. Note that such considerations ultimately extend to the reliability of any methodology: factors such as administration, vignette formulation and statistical analysis impact on the use of any experimental technique, including TVJs.

not even Chinese participants (who were generally inclined to choose the purportedly descriptivist TVJ) ever reached a proportion of TVJs "True" as high as the one in the current study: Li's (2021) Experiment 1 provides the Chinese highest proportion of TVJs "True", which peaked at 68%. While the increment of the answer "True" reinforces the discrepancy between the TVJ results and the eye-tracking ones, thereby supporting our conclusions, it is worth stressing that, even if our findings had been more in line with those of the previous literature (i.e., around 40% of answers "True"), the discrepancy between the TVJ data and the eye-tracking data would have been substantial enough for our conclusions to remain unchanged. In other words, our theses do not depend on our somewhat exceptional TVJ results.

On the other side, such a difference calls for an explanation. In our study, an eye-tracking task preceded the TVJs. Some features of the former may have impacted on the latter, such as the use of pictures and of an audio stimulus. Speculatively, if one accepts that the perspectival ambiguity for TVJs, as discussed by Domaneschi and Vignolo (2020) and Li (2021), is a real issue, one might conjecture that the use of an audio stimulus inclines participants more strongly to imagine the character uttering the sentences and thus to take her perspective.[30] The number of vignettes may also be part of the explanation. We used 10 vignettes, while almost all the previous studies used a single vignette or 4 at most (Li, 2021; Li, 2023a, Study 3). We leave this topic for further research.

[30] The use of pictures is not an absolute peculiarity of our study. In Li et al. (2018) «the stories are accompanied by clipart pictures to engage participants» (p. 109). The incorporation of pictures in Li et al.'s study was due to the inclusion of children among the participants, who may encounter more difficulties than adults in following the Gödel-Schmidt story if presented solely in a textual format.

Acknowledgements

We would like to express our gratitude to the audiences of the events where we presented this work, or parts of it, for their valuable comments and insights: the *11th European Congress of Analytic Philosophy in Vienna*, the *15th Conference of the Italian Society for Analytic Philosophy* in Alessandria, the Workshop *Experimental Semantics in Philosophy* in Trondheim, and the *XXV World Congress of Philosophy* in Rome. Special thanks to Wojciech Rostworowski for his kind invitation to the Workshop *Theoretical and Experimental Advances in Theories of Meaning and Reference* in Warsaw, and to Genoveva Martí for kindly giving us the opportunity to present an early version of this work at an online LOGOS reading group on experimental philosophy.

We are grateful to Bianca Cepollaro, Daniel Cohnitz, Michael Devitt, Jussi Haukioja, Genoveva Martí, and Wojciech Rostworowski for the enriching conversations. We extend our thanks to the two reviewers of *Linguistics and Philosophy* for their constructive and insightful feedback, which greatly improved the quality of this article.

Work on this article was supported by the PRIN project EPITHETS (Explaining Pejoratives In THeoretical and Experimental TermS, Ministero dell'Università e della Ricerca, 2022N87CR9). Nicolò D'Agruma's work on this article was first supported by a PhD scholarship from Vita-Salute San Raffaele University and later by the aforementioned PRIN project through the grant D53D23009790001 from Vita-Salute San Raffaele University.

References

Altmann, G. T. M., & Kamide, Y (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition, 73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language, 57*(4), 502–518. https://doi.org/10.1016/j.jml.2006.12.004

Arai, M., Van Gompel, R. P., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive psychology*, *54*(3), 218–250. https://doi.org/10.1016/j.cogpsych.2006.07.001

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of memory and language*, *59*(4), 457–474. https://doi.org/10.1016/j.jml.2007.09.002

Beebe, J. R., & Undercoffer, R. (2016). Individual and Cross-Cultural Differences in Semantic Intuitions: New Experimental Findings. *Journal of Cognition and Culture*, *16*(3-4), 322–357. https://doi.org/10.1163/15685373-12342182

Cappelen, H. (2012). *Philosophy without Intuitions*, Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199644865.001.0001

Chambers, C. G., & San Juan, V. (2008). Perception and presupposition in real-time language comprehension: Insights from anticipatory processing. *Cognition*, *108*(1), 26-50. https://doi.org/10.1016/j.cognition.2007.12.009

Cohnitz, D. (2015). The Metaphilosophy of Language. In J. Haukioja (Ed.), *Advances in Experimental Philosophy of Language* (pp. 85–108). Bloomsbury Publishing. https://doi.org/10.5040/9781474219815.ch-005

Cohnitz, D., & Haukioja, J. (2013). Meta-Externalism vs Meta-Internalism in The Study of Reference. *Australasian Journal of Philosophy 91*(3), 475–500. https://doi.org/10.1080/00048402.2012.717534

Cohnitz, D., & Haukioja, J. (2015). Intuitions in Philosophical Semantics. *Erkenntnis*, *80*(3), 617–641. https://doi.org/10.1007/s10670-014-9666-1

Cohnitz D., & Haukioja, J. (2021). Reference and Intuitions. In S. Biggs, & H. Geirsson (Eds.), *The Routledge Handbook of Linguistic Reference* (pp. 551–559). Routledge.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive psychology*, *6*(1), 84–107. https://doi.org/10.1016/0010-0285(74)90005-X

D'Agruma, N. (2023). From the epistemic perspectives in experimental semantics to the ambiguity of proper names: Is the inference warranted? A critical notice of Jincai Li's *The referential mechanism of proper names*. *Mind & Language*, *38*(4), 1138–1146.

Deutsch, M. (2009). Experimental Philosophy and the Theory of Reference. *Mind & Language*, *24*(4), 445–466. https://doi.org/10.1111/j.1468-0017.2009.01370.x

Deutsch, M. (2010). Intuitions, Counter-Examples, and Experimental Philosophy. *Review of Philosophy and Psychology*, *1*(3), 447–460. https://doi.org/10.1007/s13164-010-0033-0

Deutsch, M. (2015). Kripke's Gödel Case. In J. Haukioja (Ed.), *Advances in Experimental Philosophy of Language* (pp. 7–30). Bloomsbury Publishing. https://doi.org/10.5040/9781474219815.ch-002

Devitt, M. (1981). *Designation*. Columbia University Press. https://doi.org/10.7312/devi90836

Devitt, M. (2006). Ignorance of Language. Clarendon Press. https://doi.org/10.1093/0199250960.001.0001

Devitt, M. (2011). Experimental Semantics. *Philosophy and Phenomenological Research*, *82*(2), 418–435. https://doi.org/10.1111/j.1933-1592.2010.00413.x

Devitt, M. (2012a). Semantic Epistemology: Response to Machery. *THEORIA. An International Journal for Theory, History and Foundations of Science*, *27*(2), 229–233. https://doi.org/10.1387/theoria.6225

Devitt, M. (2012b). Whither Experimental Semantics? *THEORIA. An International Journal for Theory, History and Foundations of Science*, *27*(1), 5–36. https://doi.org/10.1387/theoria.3866

Devitt, M. (2015a). Testing Theories of Reference. In J. Haukioja (Ed.), *Advances in Experimental Philosophy of Language* (pp. 157–172). Bloomsbury Publishing. https://doi.org/10.5040/9781474219815.ch-003

Devitt, M. (2015b). Relying on Intuitions: Where Cappelen and Deutsch Go Wrong. *Inquiry*, 58(7-8), 669–699. https://doi.org/10.1080/0020174X.2015.1084824

Devitt, M. (2020). Stirring the Possum: Responses to the Bianchi Papers. In A. Bianchi (Ed.), *Language and Reality from a Naturalistic Perspective* (pp. 371–455). Springer. https://doi.org/10.1007/978-3-030-47641-0_19

Devitt, M. (2023). Red herrings in experimental semantics: Cultural variation and epistemic perspectives. A critical notice of Jincai Li's *The referential mechanism of proper names*. *Mind & Language*, *38*(4), 1147-1156. https://doi.org/10.1111/mila.12468

Devitt, M., & Porot, N. (2018). The Reference of Proper Names: Testing Usage and Intuitions. *Cognitive Science*, *42*(5), 155215–85. https://doi.org/10.1111/cogs.12609

Devitt, M., & Sterelny K. (1999). *Language and Reality: An Introduction to the Philosophy of Language*. Blackwell Publishers.

Domaneschi, F., & Vignolo, M. (2020). Reference and the ambiguity of truth-value judgments. *Mind & Language*, *35*(4), 440–455. https://doi.org/10.1111/mila.12254

Ferguson, H.J., & Breheny, R. (2011). Eye movements reveal the time-course of anticipating behaviour based on complex, conflicting desires. *Cognition*, *119*(2), 179–196. https://doi.org/10.1016/j.cognition.2011.01.005

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta psychologica*, *137*(2), 151–171. https://doi.org/10.1016/j.actpsy.2010.11.003

Ito, A., & Knoeferle, P. (2023). Analysing data from the psycholinguistic visual-world paradigm: Comparison of different analysis methods. *Behavior Research Methods*, *55*(7), 3461–3493. https://doi.org/10.3758/s13428-022-01969-3

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and language*, *49*(1), 133–156. https://doi.org/10.1016/S0749-596X(03)00023-8

Kripke, S. A. (1980). *Naming and Necessity*, Harvard University Press.

Li, J. (2021). The Origin of Cross-Cultural Differences in Referential Intuitions: Perspective Taking in the Gödel Case. *Journal of Semantics*, *38*(3), 415–440. https://doi.org/10.1093/jos/ffab010

Li, J. (2023a). *The Referential Mechanism of Proper Names*. Routledge. https://doi.org/10.4324/9781003047292

Li, J. (2023b). Cross-cultural variation and perspectivalism: Alignment of two red herrings?. *Mind & Language*, *38*(4), 1157–1163. https://doi.org/10.1111/mila.12470

Li, J., Liu, L., Chalmers, E., & Snedeker, J. (2018). What is in a name? The development of cross-cultural differences in referential intuitions. *Cognition*, *171*, 108–111. https://doi.org/10.1016/j.cognition.2017.10.022

Machery, E. (2011). Variation in Intuitions about Reference and Ontological Disagreements. In S. D. Hales (Ed.), *A Companion to Relativism* (pp. 118–136). Blackwell. https://doi.org/10.1002/9781444392494.ch7

Machery, E. (2012). Semantic Epistemology: A Brief Response to Devitt. *THEORIA. An International Journal for Theory, History and Foundations of Science*, *27*(2), 223–227. https://doi.org/10.1387/theoria.6223

Machery, E. (2014). What Is the Significance of the Demographic Variation in Semantic Intuitions? In E. Machery, & E. O'Neill (Eds.), *Current Controversies in Experimental Philosophy* (pp. 3–16). Routledge

Machery, E. (2017). *Philosophy Within Its Proper Bounds*. Oxford University Press. https://doi.org/10.1093/oso/9780198807520.001.0001

Machery, E. (2021). Cross-Cultural Semantics at 15. In S. Biggs, & H. Geirsson (Eds.), *The Routledge Handbook of Linguistic Reference* (pp. 535–550). Routledge.

Machery, E. (2024). Experimental Philosophy of Language: Proper Names and Predicates. In A. M. Bauer & S. Kornmesser (Eds.), *The Compact Compendium of Experimental Philosophy* (pp. 183–210). Berlin and Boston: Walter de Gruyter GmbH. https://doi.org/10.1515/9783110716931-009

Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2004). Semantics, cross-cultural style. *Cognition*, *92*(3), B1–B12. https://doi.org/10.1016/j.cognition.2003.10.003

Machery, E., Mallon, R., Nichols, S., and Stich, S. (2013). If Folk Intuitions Vary, Then What? *Philosophy and Phenomenological Research*, *86*(3), 618–635. https://doi.org/10.1111/j.1933-1592.2011.00555.x

Machery, E., Olivola, C. Y., & De Blanc M. (2009), Linguistic and metalinguistic intuitions in the philosophy of language, *Analysis*, *69*(4), 689–694. https://doi.org/10.1093/analys/anp095

Machery E., & Stich S.P. (2012). The Role of Experiment. In Fara D.G., & Russell G. (Eds.), *The Routledge Companion to the Philosophy of Language* (pp. 495–512). New York: Routledge. https://doi.org/10.4324/9780203206966-49

Magnuson, J. S. (2019). Fixations in the visual world paradigm: where, when, why? *Journal of Cultural Cognitive Science*, *3*(2), 113–139. https://doi.org/10.1007/s41809-019-00035-3

Martí, G. (2009). Against semantic multi-culturalism, *Analysis*, *69*(1), 42–48. https://doi.org/10.1093/analys/ann007

Martí, G. (2012). Empirical Data and the Theory of Reference. In W. P. Kabasenche, M. O'Rourke, & M. H. Slater (Eds.), *Reference and Referring: Topics in Contemporary Philosophy* (pp. 63–82). MIT Press. https://doi.org/10.7551/mitpress/9581.003.0004

Martí, G. (2014). Reference and Experimental Semantics. In E. Machery, & E. O'Neill (Eds.), *Current Controversies in Experimental Philosophy* (pp. 17–26). Routledge.

Martí, G. (2020). Experimental Semantics, Descriptivism and Anti-descriptivism. Should We Endorse Referential Pluralism? In A. Bianchi (Ed.), *Language and Reality from a Naturalistic Perspective* (pp. 329–41). Springer. https://doi.org/10.1007/978-3-030-47641-0_16

Maynes, J. (2015). Interpreting intuition: Experimental philosophy of language. *Philosophical Psychology*, *28*(2), 260–78. https://doi.org/10.1080/09515089.2013.815987

Nisbett, R. E. (2003). *The Geography of Thought: How Asians and Westerners Think Differently... and Why*. Free Press.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review*, *108*(2), 291–310. https://doi.org/10.1037/0033-295x.108.2.291

Norenzayan, A., Smith E., & Kim, B. (2002). Cultural preferences for formal versus intuitive reasoning. *Cognitive Science*, *26*(5), 653–684. https://doi.org/10.1207/s15516709cog2605_4

Ostertag, G. (2013). The 'Gödel' effect. *Philosophical Studies*, *166*(1), 65–82. https://doi.org/10.1007/s11098-012-0021-6

Reuter, K. (2024). The ambiguity of “true” in English, German, and Chinese. *Asian Journal of Philosophy*, *3*(1), 25. https://doi.org/10.1007/s44204-024-00150-1

Reuter, K., & Brun, G. (2022). Empirical Studies on *Truth* and the Project of Re-engineering *Truth*. *Pacific Philosophical Quarterly*, *103*(3), 493–517. https://doi.org/10.1111/papq.12370

Ronderos, C. R., Guerra, E., & Knoeferle, P. (2023). When sequence matters: The processing of contextually biased German verb–object metaphors. *Language and Cognition: An Interdisciplinary Journal of Language and Cognitive Science*, *15*(1), 1–28. https://doi.org/10.1017/langcog.2022.22

Ronderos, C. R., & Domaneschi, F. (2023). Anticipating the *Damn* Referent: How Comprehenders Rapidly Retrieve the Speaker's Attitude When Processing Negative Expressive Adjectives. *Cognitive Science*, *47*(5), e13295. https://doi.org/10.1111/cogs.13295

Sytsma, J., & Livengood, J. (2011). A New Perspective Concerning Experiments on Semantic Intuitions, *Australasian Journal of Philosophy*, *89*(2), 315–32. https://doi.org/10.1080/00048401003639832

Sytsma, J., Livengood, J., Sato, R., & Oguchi, M. (2015). Reference in the Land of the Rising Sun: A Cross-cultural Study on the Reference of Proper Names. *Review of Philosophy and Psychology*, *6*(2), 213–230. https://doi.org/10.1007/s13164-014-0206-3

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. G. (2000). Eye Movements and Lexical Access in Spoken-Language Comprehension: Evaluating a Linking Hypothesis between Fixations and Linguistic Processing. *Journal of Psycholinguistic Research*, *29*, 557−580. https://doi.org/10.1023/A:1026464108329

Tanenhaus, M. K., & Spivey-Knowlton, M.J. (1996). Eye-Tracking. In Grosjean, F, & Frauenfelder U. (Eds.), *Language and Cognitive Processes: A Guide to Spoken Word Recognition Paradigms* (Vol. 11, pp. 583–588). Psychology Press. https://doi.org/10.1080/016909696386971

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. https://doi.org/10.1126/science.7777863

van Dongen, N., Colombo, M., Romero, F., & Sprenger, J. (2021). Intuitions About the Reference of Proper Names: a Meta-Analysis. *Review of Philosophy and Psychology*, *12*(4), 745–774. https://doi.org/10.1007/s13164-020-00503-8

Vignolo, M. & Domaneschi, F. (2018). Referential intuitions are still problematic. *Analysis*, *78*(3), 472–483. https://doi.org/10.1093/analys/anx148