Experimentally Testing the Intuitions about Semantic Reference

Ana Puljić (s1352997@sms.ed.ac.uk)
Leonidas A. A. Doumas (alex.doumas@ed.ac.uk)
Department of Psychology, the University of Edinburgh,
7 George Square, Edinburgh EH8 9JZ

Abstract

The debate about semantic reference between Frege's (1948) descriptivism and Kripke's (1972) causal theory of reference has recently been approached through experimental psychology. However, no consensus has been reached on the direction of the results. While some studies face clear methodological charges, even those that are currently uncontested do not reach a mutual conclusion. We propose a novel experimental paradigm with methodology designed to evade the problems of previous studies. Contrary to the past literature, we find a prevalence of descriptivists under lenient criteria for consistency across trials, while under strict criteria we find an equal amount of descriptivists and hybrids, with low numbers of referentialists (causal theory of reference) under both criteria. We suggest an interpretation of this result, and where future research might head.

Keywords: proper names; descriptivist theory; causal theory of reference; semantic reference

1. Introducing the Reference Debate

A central issue in the philosophy of language is the debate between descriptivism and the causal theory of reference regarding the meaning and reference of proper names. For descriptivists the description constitutes the meaning of a name and thereby picks out the referent. For causal theorists a description only fixes the name in the initial baptism, after which the name is passed on through a causal chain of communication. Kripke's (1972) widely accepted refutation of descriptivism relies on thought experiments which demonstrate how descriptivism reaches wrong reference judgements. As thought experiments rely on assuming the generality of the suggested intuitions, some experimental philosophers have decided to test whether Kripke's assumptions hold with non-philosophers as well. Before introducing such studies, we outline the tension between Frege's (1948) descriptivism and Kripke's causal theory through Kripke's objections to cluster descriptivism.

To advance a point against Frege's descriptivism, Kripke (1972) differentiates between two roles that descriptions can have with respect to the reference of proper names: being synonymous with them or fixing their referent. Kripke suggests names should be viewed as rigid designators: their reference will remain the same across possible worlds by virtue of referring to the objects directly, instead of being synonymous with their descriptions. Hence names lack meaning additional to picking out the referent. In contrast, Frege differentiates between a *reference* of the proper name as the object it designates, and the *sense* of the name as the mode of presentation of the object, often understood as its associated descriptive content. Consequently, Kripke indicates that for Frege proper names are actually abbreviated descriptions as they have an associated sense (e.g., *Aristotle exists* is read as *there was a man who did X*). The referent of the name is then picked out through this description. Such non-essentialist descriptions are, however, merely contingent facts and therefore, their reference will change across possible worlds (i.e., there might have been some other man who did things accredited to Aristotle, making him 'Aristotle'). Instead, Kripke suggests people use some accidental descriptive property of an individual to fix the reference in the initial baptism, which then fixes the reference across all possible worlds. Kripke's theory allows this description to later turn out false, as it does not constitute the meaning of the name. Names are then passed along from speaker to speaker, where the receiving party must intend to use the same reference as the giving party.

Kripke later introduces and formalizes Searle's (1958) cluster descriptivism as the strongest version of simple descriptivism. Cluster descriptivism changes the focus from a single property to a whole cluster of properties. He breaks it down into multiple theses, all of which he rejects, concluding that the theory fails. As most experimental philosophy studies focused on the third thesis, we will outline Kripke's argument against this thesis only.

The thesis states that if most properties are satisfied by one unique object X, then X is the referent of the name N (p.71). Against this assertion, Kripke provides a thought experiment as a counterexample: unbeknownst to the wider public, it was Schmidt and not Gödel who proved the incompleteness of arithmetic – a description usually associated with the name 'Gödel'. When we then say 'Gödel', seeing as Schmidt satisfies the description commonly associated with 'Gödel', cluster descriptivists will then have to claim that we are actually referring to *Schmidt*. Kripke suggests this is false as surely by 'Gödel' we are just referring to *Gödel*. Hence, the thesis seems false.

2. Previous Experimental Studies on Reference

As demonstrated, a large part of Kripke's refutation of descriptivism rests upon thought experiments intended to support causal theory intuitions. Consequently, some studies within experimental philosophy have attempted to verify the legitimacy of these conclusions by testing the intuitions of non-philosophers on modified Gödel cases. In this section, we briefly outline and assess some of these studies.

The first experimental study to engage with the reference debate was Machery, Mallon, Nichols and Stich's (2004) study, whose vignettes were based on Kripke's Gödel case, phrased in the third-person view. Participants had to answer

whether a speaker who uses the name 'Gödel' is talking about the discoverer of the incompleteness theorem or the person who got credited for it. Results indicated that Westerners gave more causal answers while East Asians gave more descriptivist answers. The authors took this to suggest that philosophers should either avoid assuming universality of their semantic intuitions or that they should offer an argument for why their intuitions should be superior to those of the average speaker. Genone and Lombrozo's (2012) study confirmed these results and extended it from people proper names to natural and nominal kind concepts on which the philosophical theories focus only marginally. Furthermore, they also found a within-participants difference, suggesting individuals are not consistently pure descriptivists or referentialists.

These two studies have been criticized for potentially testing the speaker's reference (what the speaker intended to refer to) instead of the semantic reference (what the name itself semantically refers to), while only the latter is relevant for a theory of semantic reference (Devitt, 2011; Martí, 2009). In addition, the said studies seem to investigate *meta-linguistic* intuitions about what the term *refers to*. Martí (2009) suggests experiments should focus on *linguistic* intuitions about how a term should be *used* instead.

Li, Liu, Chalmers and Snedeker (2018) confirmed Machery et al.'s (2004) results from a developmental perspective, suggesting the differentiation in reference intuitions happens early on. Namely, they found that both children and adults from East Asia are more descriptivist than their Western counterparts. The authors hypothesized this might be due to culturally different perspective-taking tendencies. It has been found that Westerners adopt the perspective of the speaker less readily than East Asians do. Hence, Westerners might employ the perspective of the omniscient reader in the third-person vignettes and judge the truth relative to all the information available to themselves. Meanwhile, East Asians may adopt the perspective of the speaker within the vignette and thus judge the referent relative to the description available to the speaker. The authors nonetheless acknowledge the above problems regarding speaker's reference for their own design as well.

These concerns are addressed by the study of Domaneschi, Vignolo, and Di Paola (2017) who managed to experimentally distinguish whether the participants' answers were guided by the speaker's or the semantic reference. Their results supported the causal theory of reference, however, they also found an unexpected difference between people names and geographical names.

Domaneschi et al. (2017) also received indirect support from Jylkka, Railo and Haukioja's (2009) experiment which indicated participants tended to exhibit more semantically externalist intuitions than internalist and hybrid intuitions for natural kind concepts. Semantic externalism[1] is usually taken to be implied by the causal theory of reference, which thus aligns these results with the findings of Domaneschi et al. (2017). A recent paper by Tobia, Newman, and Knobe (2017), however, indicated that participants reported intuitions based on *both* externalist *and* internalist properties for natural kinds, seemingly supporting the hybrid externalism that Jylkka et al. (2009) rejected. Similarly, Nichols, Pinillos and Mallon (2016) suggested an ambiguity theory of reference for natural kind terms based on their results. Namely, depending on the context, a natural kind term sometimes referred through a descriptive mode of reference and sometimes through a causal one. These results are also compatible with Genone and Lombrozo's (2012) findings, in that they also suggest both descriptive and causal mechanisms contribute to categorization.

However, the latter two studies—along with Machery et al. (2004), Li et al. (2018), and Tobia et al. (2017), as those suggesting an indeterminacy between the two theories—disagree with the findings of Jylkka et al. (2009) and Domaneschi et al. (2017) whose results support the causal theory of reference. While the latter two do not check for consistency within individuals like Genone and Lombrozo (2012), their results seem well-founded as they account for the aforementioned methodological problems of Machery et al. (2004), Genone and Lombrozo (2012), and Li et al. (2018). However, this is also seemingly true of Nichols et al. (2016) and Tobia et al. (2017) whose results disagree with the former two. Hence, even studies with prima facie no methodological issues reach different conclusions.

In light of this stalemate, we propose a novel paradigm for testing reference intuitions that addresses the methodological flaws of the existing literature. In a basic vignette, two item names are introduced with two pictures representing people, fictitious places or fictitious natural kinds. One of them is assigned a trivial or uniquely identifying descriptive property which, at the end of the vignette, gets mismatched with the wrong picture while the name is left out. The participants are then asked to determine the name of the item so presented. Referentialists should choose the picture-based name while descriptivists should choose the property-based name. Half of the trials also include an additional step of negation before the prompt at the end in which the property is negated for the first item and re-introduced with the other item.

This setup addresses possible methodological shortcomings of previous literature in three ways. First, instead of using third-person reference judgements that potentially test meta-linguistic intuitions, this paradigm tests the linguistic intuitions of the participants directly through employing first-person requests for assigning referents.

Second, the process is reversed as the critical prompt starts from the referent which is represented by a picture, after which the participants choose its name. Here, participants' intuitions should only be led by the semantic reference. The possibility of the speaker's reference driving the answers is eliminated by reversing the usual paradigm where the name's referent has to be determined from the name itself. Additionally, a first-person vignette excludes the existence of a third speaker about whose intentions participants could (mistakenly) cognize.

[1] Factors external to the speaker determine a term's meaning.

Third, the effects of different noun types found in the literature so far, namely people proper names, natural kinds, and geographical names, as well as a newly added variable of uniqueness of the associated descriptive property and a variable of negation were examined both between and within participants across multiple trials. These were included as potential contextual factors which might influence the saliency of different referential mechanisms. While there are theoretical reasons for believing uniquely identifying properties might make descriptivist intuitions stronger, theories remain mostly agnostic in regards to item types other than proper nouns. An effect of item type would hence likely be attributed to some pragmatic factor. Similarly as for item types, the inclusion of explicit negation (trial type) is not theory-driven, but it serves to investigate the possibility that, in addition to the newly corrected association, participants might still retain the old association of a name even after it is negated. An effect of trial type could potentially indicate participants failed to experience a mismatch in trials containing the negation, as the old association they retained made the last picture-property pairing match. This would skew the results to be seemingly more picture based and hence potentially falsely causal.

Generally, a referentially heterogeneous population would suggest the public is not uniformly descriptivist or causal and thus entertains intuitions different to philosophers. A majority of internally inconsistent participants would indicate that people perhaps use a combination of the two theories. In that case, a hybrid theory of reference may hold, where different contextual factors trigger the opposite reference mechanisms. All of these would question some of Kripke's arguments and indicate the need to identify the contextual factors behind the different intuitions.

3. Experiment

This experiment had three main aims. The first and the primary aim was to examine whether the population is homogeneous in reference intuitions. The second aim was to determine whether participants had internally consistent intuitions across trials. The third aim was to investigate whether different property types, item types and trial types had explanatory effects on reference intuitions both across and within participants.

3.1. Method

3.1.1. Participants Forty-four participants completed the experiment via Mechanical Turk. After checking the control trials for comprehension checks, 4 participants failed to satisfy the preset 75% accuracy requirement and were removed from the analysis. Mean age was 37.15 (*SD* = 10.66). Out of the remaining 40 participants, 36 were American and 4 were Indian, with 20 of them having completed higher education, 9 having completed some university and 11 having completed only high school. As the sample was 90% Western, it was assumed nationality was not explaining the variation within the results.

3.1.2. List of Stimuli Items used in the experiment belonged to one of the three domains covered within existing research: people names, natural kinds, and geographical names. The latter two were fictitious to control for the participants' existing knowledge of features of real natural kinds and geographical places. Properties that were crossed with the items were either uniquely identifying (applies to only one individual) or non-uniquely identifying (applies to many). As these item-property combinations were further combined with two critical trial types (A and B) and two parallel control trial types, all of which are discussed below, the resulting formula of our fully crossed design yielded 3 item types crossed with 2 property types which are further crossed with 4 trial types. The total number of the minimum of fully crossed combinations of all three variables was hence 24, where there was exactly one crossing of each trial type with each level of item type and each level of property type[2]. Each participant underwent the said minimum of trial combinations, which yielded 24 trials per participant, where half were critical and half were control trials. We chose to start with only 24 trials due to the very poor cost-efficiency, where doubling the number of trials per participant would only result in one additional crossing between each level of the three variables. This limitation is discussed further in section 4.

3.1.3. Design and Procedure The experiment was ethically approved (code 24-1718/1) by the Psychology Research Ethics Committee at the University of Edinburgh**.** In the experiment, participants underwent a series of trials, each constituted of a vignette and a prompt (Figure 1). Participants could see all the steps of one trial at once, and they moved between trials by pressing a key that indicated their response. In a basic vignette, participants were shown two pictures of two items paired with their names, *N1* and *N2*. A specific descriptive property, *P*, that was either unique or non-unique, was then introduced as belonging to the first item. Two trial types diverged at this point in regards to the existence of explicit negation of property *P*. For *B-type* trials it was revealed that due to some epistemic mistake, property *P* actually belongs to the second item, explicitly negating the property for the first item. For *A-type* trials, this negation was left out. Lastly, a picture of the item which was not last associated with the property was shown without its name, with the suggestion that this item has the property *P*. This created an item-property mismatch. The participants were then asked to determine whether the item from the last picture was *N1* or *N2*.

Now, here is an outline of a top-to-bottom screen of a B-type critical trial with human item type and unique property type: (1) a photo of a woman and *"This is Dora",* (2) a photo of another woman and *"This is Jenia",* (3) the first photo and *"Dora is the best speller under the age of 24"*, (4) *"Actually, that was false. This no longer holds as this person turned 25 yesterday",* (5) the second photo and

[2] See https://tinyurl.com/ycsx8zbm for the full list of item pairs, properties, photos, and negation explanations used.

"Actually, Jenia is the best speller under the age of 24", and (6) the first photo and *"This person is the best speller under the age of 24. Who is this, Dora or Jenia?"* In an A-type trial, steps (4) and (5) would be missing and in step (6) the second photo would be shown instead.

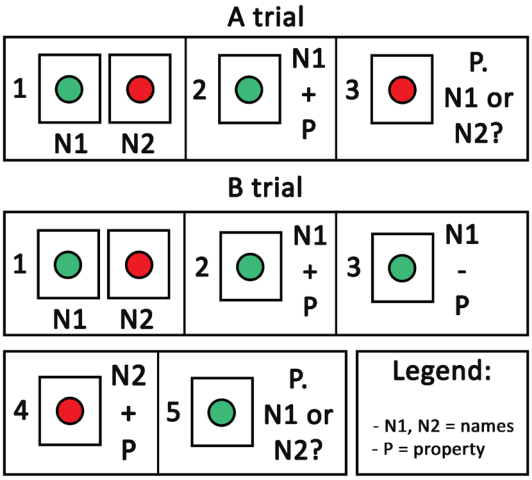

Figure 1: Graphical representation of critical trial types.

Choosing the item name which was last associated with the property was assumed to privilege the descriptive property as an identifier, which suggests descriptivist intuitions. Choosing the picture-based item name, in contrast, focused on the causal history of the name first associated with the item, suggesting causal intuitions. All participants also underwent control trials which contained no item-property mismatch, as the last picture depicted the last item associated with the property. These had objectively correct answers and were used as comprehension checks.

Our reasoning about what makes an answer descriptivist or causal followed the following theoretical extrapolations. According to the causal theory, the truth of a name's associated descriptions should not dictate whom the name refers to past the initial introduction. Namely, in baptisms by description, the property merely fixes the referent of a name which then rigidly refers to its referent without the mediation of the property. For descriptivism, however, people's choices are led by the name's associated property as it constitutes the name's meaning and picks out the referent. This should hold even if the *picture* associated with the name possesses *a number* of properties and people associate some of *these* properties with the name, *in addition* to the single *introduced* property. At least when the introduced property is *uniquely identifying*, it must be central to the meaning of the name for descriptivists, as such a property is seemingly essential of the named individual. Consequently, descriptivists should choose the names based off unique properties. However, arguably even a mismatched *non-unique* property would lead descriptivists to choose the property-based name. This is because it remains the only non-picture information that the participants know about the name it was introduced with. Hence, if this assumption holds for descriptivists, a non-unique property should also contribute to the name's meaning. Otherwise, some descriptivists will choose a picture-based name when the property is non-unique, thus making the results more falsely causal. This will be accounted for in a follow-up discussed in section 4. For the purposes of the present experiment, it will be presumed that the said assumption holds and descriptivists should choose the names based off both unique and non-unique properties.

3.2. Results

For the contextual effects of item type, property type and trial type that we hypothesized might influence reference intuitions, we built a maximal and a minimal mixed effects logistic regression model using the critical trials data, with name choice (descriptivist vs. causal) as the dependent binary variable. Both models included participants as random effects, while the maximal model also had all the independent variables (item type, property type, and trial type) and interactions between them as fixed effects. Comparison of the two models via ANOVA suggested the two models were not significantly different (X^2 (11, N = 40) = 10.11, p>.05). Hence, the additional variables of the maximal model did not add to the explanation of the variance. That is, participants' reference judgements were unaffected by different types of items, properties and trials.

Further, for examining the within-participant consistency and differences across participants, chi-square tests were used after the participants had been tagged as either consistently causal, consistently descriptivist or hybrid. Since there is no objective criteria for how many trials out of 12 critical trials one should answer in a uniform way to be considered consistent, the tagging was done over five instances. The lower cut-off point of the number of uniform answers for being tagged as consistent ranged from having a minimum of 7 (very lenient criteria, 58%) to a minimum of 11 (fairly strict criteria, 92%) out of 12 in total.

When looking at the within-participant consistency, participants tagged as consistently causal or consistently descriptivist represented the consistent group while the ones tagged as hybrid represented the inconsistent group. The two were compared via chi-square for the different criteria. For 7+, 8+ and 9+ criteria, the results indicated that there were more consistent than inconsistent participants, while for the 10+ and 11+ criteria, there was no difference in the number of consistent and inconsistent participants (Table 1).

Table 1: Frequencies of Consistent Participants (N = 40) and Chi-Squares against Inconsistent Participants

Criteria	7+	8+	9+	10+	11+
Frequency	38	33	31	26	16
X^2 (1, N = 40)	32.40*	16.90*	12.10*	3.60	1.60

Note. *p < .05

Regarding the between-participants differences, when all three groups were compared to each other via the chi-square test of goodness-of-fit for the different criteria, all significantly differed from an equally distributed population

across all the criteria (see Table 2, Figure 2). For the 7+ criterion, descriptivists significantly outnumbered referentialists (X^2 (1, N = 40) = 18.10, p<.05) and hybrids (X^2 (1, N = 40) = 55.97, p<.05), while referentialists outnumbered hybrids (X^2 (1, N = 40) = 3.79, p<.05). For the 8+ criterion, there was an equal amount of referentialists and hybrids, both of whom were outnumbered by descriptivists (X^2 (1, N = 40) = 16.71, p<.05). The same pattern was true of the 9+ criterion, where there was no significant difference between the amount of referentialists and hybrids (X^2 (1, N = 40) = 0.78, p>.05), while the number of descriptivists was significantly higher than both (X^2 (1, N = 40) = 21.07, p<.05; X^2 (1, N = 40) = 13.00, p<.05, respectively). For the 10+ criterion, there was no significant difference between the number of descriptivists and hybrids (X^2 (1, N = 40) = 3.22, p>.05), both of whom significantly exceeded referentialists in number (X^2 (1, N = 40) = 20.57, p<.05; $X2$ (1, N = 40) = 7.47, p<.05). A reversal happened for the 11+ criterion, where there were more hybrids than descriptivists (X^2 (1, N = 40) = 4.06, p<.05) and referentialists (X^2 (1, N = 40) = 25.13, p<.05), as well as statistically more descriptivists than referentialists (X^2 (1, N = 40) = 9.45, p<.05).

Table 2: Chi-Squares for Between-Participant Frequencies for Different Criteria

Criteria	Des.	Cau.	Hyb.	X^2 (2, N = 40)
7+	29	9	2	29.45*
8+	26	7	7	18.05*
9+	26	5	9	18.65*
10+	23	3	14	15.05*
11+	14	2	24	18.20*

Note. *p < .05.

Overall, participants were internally consistent up until the 10+ criteria, after which the number of consistent participants equates with the number of inconsistent participants. In regards to the between-participants difference, from 7+ to 9+ criteria there were more descriptivists than referentialists and hybrids. At 10+ criteria there is an even spread of descriptivists and hybrids which outnumber the very small number of referentialists, while after 11+ criteria hybrids take the lead over descriptivists.

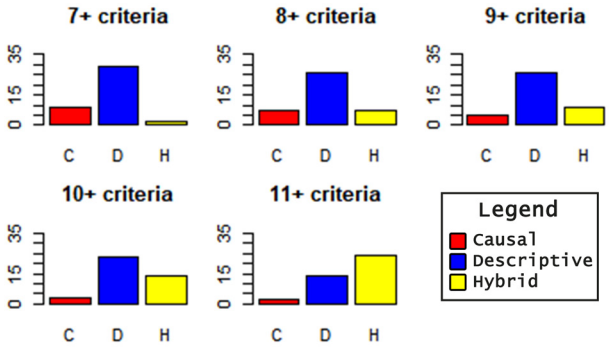

Figure 2: Frequencies of between-participants differences in reference intuitions across the consistency criteria.

4. General Discussion

There seems to be no single conclusive way to make sense of the findings of this study without choosing a preferred criteria first. We can nonetheless note that neither the most lenient nor the most stringent criteria align the results with Kripke's theory. Moreover, the results indicate a trend in the opposite direction while the criteria is lenient, with the population seeming primarily descriptivist with only a few scarce referentialists. As the criterion gets stricter, hybrids gradually outnumber the descriptivists while referentialists remain almost non-existent. This is in contrast to the past literature which mostly reported equally heterogeneous population in regards to descriptivists and referentialists, with only two studies reporting evidence for a primarily homogenous population, and both of these results were in favor of the causal theory of reference. Current results can either be taken to suggest that the population is primarily descriptivist, or that a large portion of the population uses both theories but employs each of them in different contexts. Both of these interpretations allow us to conclude that the population does not uniformly share Kripke's intuitions regarding the Gödel-like cases.

However, rather than interpreting our results as indicating that the population is primarily descriptivist, it seems such results might instead be inviting us to take the conclusions of this literature with a grain of salt. Namely, it becomes quite plausible to consider that the phenomenon under investigation is either sufficiently unstable to vary from paradigm to paradigm, or even theoretically poorly operationalized. While one could always combat this by saying the paradigms are at fault here – that is, they are the ones failing to operationalize the phenomenon correctly – it nonetheless seems equally plausible to speculate whether the phenomenon *itself* is accurately captured by either of the two proposed theories. Perhaps neither of the two philosophical theories adequately describe and explain how *real* people use *natural* language, or specifically proper names within the natural language, in day to day life.

The possibility remains that the solution lies somewhere in between, in the combination of the two theories which are evoked at different times in different contexts. This stays a valid direction for future research to take in which the search for the contextual determiners continues. The present study has contributed to that cause, albeit by providing an evidence of absence, as contrary to our belief, none of the contextual factors we investigated affected the reference intuitions. There remain a few other factors we wish to examine in the future, whose effects, or the lack of thereof, might indicate whether this route should ultimately be relinquished. These are briefly discussed below.

The next follow-up experiment would focus on varying the number of assigned properties and the amount of them that is negated. This will account for the possibility that participants have multiple associated properties, instead of just the explicitly introduced property, due to the name's associated picture. This should show an effect for descriptivists while it should not show an effect for

referentialists. Namely, descriptivism would predict *multiple* associated properties would always be favored in picking out an individual over the *single* associated property[3]. If there is an effect of this manipulation, then it might be speculated that in the current study there was an underestimate of descriptivist answers. In that case, some covert-descriptivists might have chosen the name based off the picture when the property was not unique, as they believed the picture retained more of the name's associated properties in comparison to the one explicit property that was mismatched and not unique. This could potentially explain away some of the apparent hybrids or even the little referentialists that were found in the present experiment.

Additionally, strength of conviction could be added for each response in our present design to examine whether certain intuitions relate to more or less certain participants. A measure of one's readiness to accept under-informative implicatures (e.g., Antoniou, Cummins, & Katsos, 2016) could also be included, as our vignettes are arguably fairly ambiguous. Consequently, the results rely on the participants' gratuity towards the ambiguity of the vignettes and the assumption that they do not mistrust the setting.

Future research could also focus on compiling the different experimental paradigms from the literature and then test them as a within-participants variable. Doing so would allow for a comparison of paradigms that have so far found contrasting results. With that, instead of speculating, we could begin the discussion about correct operationalization of the phenomenon at hand in virtue of both the experimental paradigms and the theories behind it.

Lastly, it is important to note the limitations of the small sample size used in this study, as well as of the small number of trials per participant which ensured that a crossing between each level of each variable occurs only once. Despite that, our pilot[4] with a sample size of N = 115 yielded the same pattern of results as we did now due to which we decided to go with a smaller sample size. We are nonetheless in the process of running this design on a bigger sample with a larger number of trials and added strength of conviction per trial. We hope to validate and expand on the current conclusions in the overpowered sample.

In summary, this study presented a novel experimental paradigm for inquiring into the question of semantic reference within the philosophy of language. The interpretation of the results of the present experiment depend on the chosen consistency criteria. If we take the more lenient consistency criteria, then it seems the results indicate a fairly strong prevalence of descriptivists within the population, aside a comparably small portion of hybrids and referentialists. For the lenient criteria, participants appeared to be predominantly consistent across trials. The results are interpreted somewhat differently when adhering to the stricter criteria according to which the number of hybrids is equal to or greater than the number of descriptivists. This ultimately leads to an equal amount of internally consistent and inconsistent participants. For both criteria, the number of causal participants remains very low when compared to the number of descriptivists and hybrids. These findings, especially when read through the lenient criteria, go contrary to the trends found in the past literature. These report an equal amount of descriptivists and referentialists, or alternatively, the predominance of referentialists. While some might take these results to motivate a theoretical move towards descriptivism, these could also be taken to suggest that the phenomenon of semantic reference is either a variable phenomenon which is sometimes descriptivist and sometimes causal, or alternatively, that these two theories have simply failed to capture the true way in which people use proper names.

[3] Based on the third thesis where a referent of a name satisfies the *most* of the name's associated properties.

[4] In the pilot, one trial was spread across multiple screens, and the explanations of negations for B-type trials were less elaborate.

Acknowledgments

We are grateful to Dr. Hugh Rabagliati for insightful discussions and comments on the work. We also thank the anonymous reviewers for their constructive guidance.

References

Antoniou, K., Cummins, C., & Katsos, N. (2016). Why only some adults reject under-informative utterances. *Journal of Pragmatics, 99*, 78-95.

Devitt, M. (2011). Experimental semantics. *Philosophy and Phenomenological Research, 82*(2), 418-435.

Domaneschi, F., Vignolo, M., & Di Paola, S. (2017). Testing the causal theory of reference. *Cognition, 161*, 1-9.

Frege, G. (1948). Sense and reference. *The philosophical review, 57*(3), 209-230.

Genone, J., & Lombrozo, T. (2012). Concept possession, experimental semantics, and hybrid theories of reference. *Philosophical Psychology, 25*(5), 717-742.

Jylkkä, J., Railo, H., & Haukioja, J. (2009). Psychological essentialism and semantic externalism: Evidence for externalism in lay speakers' language use. *Philosophical Psychology, 22*(1), 37-60.

Kripke, S. (1972). Naming and Necessity. In D. Donald & H. Gilbert (Eds.), *Semantics of Natural Language*. Dordrecht: Reidel: 253-355, 763-769.

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.

Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2004). Semantics, cross-cultural style. *Cognition, 92*(3), B1-B12.

Martí, G. (2009). *Against semantic multi-culturalism. Analysis, 69*(1), 42-48.

Nichols, S., Pinillos, N. Á., & Mallon, R. (2016). Ambiguous reference. *Mind, 125*(497), 145-175.

Searle, J. (1958). Proper Names. *Mind, 67*(266), 166–73.

Tobia, K. P., Newman, G. E., & Knobe, J. (2017). *Water Is and Is Not H2O.* Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2933100