On Experience-Driven Semantic Judgments: A Case Study on the Oneiric Reference Constraint

Jeruen E. Dery (dery@zas.gwz-berlin.de)
Hazel Pearson (pearson@zas.gwz-berlin.de)
Zentrum für Allgemeine Sprachwissenschaft
Schützenstraße 18, Berlin, Germany 10117

***Abstract*—Using the interpretation of pronouns in dream and belief reports as a test case, we show that semantic judgments can vary as a function of experience. We present findings from three studies where semantic judgments for pronoun interpretations were affected by a) repeated exposure, or b) the experimental task which elicited the judgments. Our findings stress the importance of converging evidence from multiple tasks and paradigms when testing and formulating theoretical hypotheses.**

I. Introduction

Previous research has shown that some syntactic and semantic judgments are affected by the amount of exposure/experience one has with these structures. Syntactic and semantic acceptability can either decrease or increase as a function of amount of experience. For example, repeated exposure to the same words or phrases causes listeners to temporarily lose its meaning, a phenomenon known as "semantic satiation" [1], [2]. Conversely, the phenomenon sometimes referred to as "linguist's disease" or "syntactic satiation", first reported in [3], refers to the observation that certain syntactic structures that were initially judged as ungrammatical slowly sound grammatical over time [4], [5].

As the literature on syntactic satiation shows, investigating whether certain linguistic phenomena exhibit a repeated exposure effect or not is important, because the presence or absence of a repeated exposure effect typically has implications for theoretical assumptions regarding the architecture of the grammar [6]. The linking assumption here is that structures that are unacceptable due to a constraint imposed by the grammar should not exhibit repeated exposure effects.[1] The repeated exposure effect is therefore a litmus test when determining whether the (un)acceptability of a particular structure is due to a grammatical constraint or not. In the current study, we explore the repeated exposure effect in the context of semantic judgments on pronoun interpretations in dream and belief reports.

In the discourse in (1), there are four potential readings of the pronouns in bold, listed in (2).

(1) There were two authors, Carol and Sandra. Carol dreamed that she was Sandra and **she** was buying **her** book.

(2) a. Sandra buys Sandra's book. (*de se* + *de se*)
b. Carol buys Sandra's book. (*de re* + *de se*)
c. Sandra buys Carol's book. (*de se* + *de re*)
d. Carol buys Carol's book. (*de re* + *de re*)

It is claimed, however, that only three of these readings are attested; the *de re* + *de se* reading is claimed to be ruled out by the so-called 'Oneiric Reference Constraint' (ORC) which prohibits the c-command of a *de se* pronoun by a corresponding *de re* pronoun [7]. The ORC is taken to follow from general syntactic constraints that prevent a binding configuration from being established in the presence of an intervening element – in this case, the *de re* element. The mechanism underlying the relevant semantic judgments is thus at root a syntactic one.

Experimental evidence for the ORC is claimed to have been given in [8]. An experimental paradigm using a novel two-picture forced-choice task was designed to test the acceptability of the *de re* + *de se* reading by pitting a picture depicting it against a picture depicting the *de re* + *de re* reading. The items described scenarios where a *de re* + *de re* reading would be possible but unlikely (e.g., *an author buying her own book*). The results showed that while participants preferred a *de se* + *de re* reading over a *de re* + *de re* reading, participants reliably preferred the *de re* + *de re* reading over the *de re* + *de se* reading. This result is interpreted by [8] as evidence that the *de re* + *de se* reading is indeed ruled out by the grammar, as claimed by [7].

However, we argue that the situation is more complex than it seems. We have reason to question whether the ORC indeed holds. In three studies, we show that participants' judgments can be modulated as a function of experience. In Study 1, we reanalyze Experiment 2 reported in [8] and show that participants reject the *de re* + *de se* reading only after a considerable amount of exposure. In Study 2, using the same paradigm, we test the ORC with a different verb with counterfactual semantics and demonstrate that participants initially accept the *de re* + *de se* reading but slowly reject it with increased exposure. Finally in Study 3, we utilize a single-picture rating task, where participants are not comparing the appropriateness of two readings simultaneously and choosing one over the other. In this single-picture paradigm, participants reliably preferred the *de re* + *de se* reading over the *de se* + *de re* reading.

[1] Previous studies have used the term "satiation" irrespective of whether linguistic acceptability improved (as in the case of syntactic satiation) or deteriorated (as in the case of semantic satiation). Because the term "satiation" seems to imply a particular direction of change of linguistic acceptability, we refrain from using this term for the remainder of this paper.

II. Study 1: Is the ORC Effect consistent over time?

We reanalyzed Experiment 2 of our previous study [8], which demonstrated that comprehenders preferred a *de re* + *de re* reading (which describes a possible but unlikely scenario) over a *de re* + *de se* reading (which is supposedly ruled out by the ORC) when forced to choose between the two. The purpose of the reanalysis is to investigate whether participants' judgments were stable during the course of the experiment or not. If participants reliably reject pictures depicting the *de re* + *de se* reading throughout the experiment, then this would constitute strong evidence that this reading is ruled out by the grammar as claimed by the ORC. If on the other hand, participants' judgments change as a function of exposure, then this would raise the possibility that some other factor is responsible for the observed effect.

A. Method and Results of Reanalysis

Unlike in [8], we analyzed participants' semantic judgments using a binary logistic regression model with Comparison Type, presentation order, and their interaction, as predictors.[2] Comparison Type is a predictor that has two values, Comparison Type A (*de re* + *de re* vs. *de re* + *de se*) and Comparison Type B (*de re* + *de re* vs. *de se* + *de re*), depending on which pair of pictures participants were presented with. We allowed the model to be adjusted by items, subjects, and lists, in order to account for random effects. A significant effect of Comparison Type was observed ($z = -6.26, p < 0.0001$), suggesting that the pattern of responses differed depending on which pair of readings participants were presented with. We also observed a significant effect of presentation order ($z = 2.48, p = 0.01$), indicating that the proportion of responses changed as a function of exposure. There was no significant interaction between Comparison Type and presentation order ($z = -1.7, p = 0.07$), suggesting that the pattern of responses changed over time irrespective of which pair of readings participants were evaluating.

We then applied the split-half method to investigate whether semantic judgments were consistent throughout the course of the experiment [9], [10]. The experiment was originally conducted via Amazon Mechanical Turk [11], and contained 18 experimental items interspersed among 54 fillers. Since participants had the chance to terminate the experiment at any time, not all participants saw all 18 experimental items. Hence, we divided the dataset into two halves, where the "first half" consisted of experimental items that had Presentation Orders 1 to 4 ($n = 349$), and the "second half" consisted of experimental items that had Presentation Orders 5 to 18 ($n = 305$). As in [8], we then analyzed both halves using a binary logistic regression model with Comparison Type as a predictor. On both halves, there was a significant effect of Comparison Type (for the first half, $z = -7.38, p < 0.0001$; for the second half, $z = -8.97, p < 0.0001$): the pattern of judgments was different in both halves when the *de re* + *de re* reading was being compared against the *de re* + *de se* reading, as opposed to against the *de se* + *de re* reading.

[2]We thank the reviewers for the suggestion to incorporate presentation order in our analyses.

TABLE I. Proportions of semantic acceptability across first and second halves for Experiment 2 in [8]

First Half	*de re+de re*	*de re+de se*	*de se+de re*	*p-value*
Comparison A	48%	51.7%		0.69
Comparison B	9%		90.9%	<0.0001
Second Half				
Comparison A	64.5%	35.4%		<0.0001
Comparison B	7.4%		92.5%	<0.0001
Overall				
Comparison A	56%	43%		0.02
Comparison B	8%		91%	<0.0001

Next, we conducted pairwise T-tests within comparison type on both halves in order to see whether there were significant differences in the proportion of semantic judgments. Table I shows the distribution of proportions within Comparison Types A and B on both halves, as well as for the overall experiment. For the first half, the proportion of responses in Comparison Type A was not significantly different ($t = 0.4, p = 0.69$): participants did not exhibit a preference for one reading over the other. On the other hand, the proportion of responses in Comparison Type A for the second half was significantly different ($t = -5.26, p < 0.0001$). This suggests that participants only systematically rejected the *de re* + *de se* reading (preferring instead the *de re* + *de re* reading, which describes a possible but unlikely scenario, see 2d above) after encountering this structure multiple times. The proportion of responses in Comparison Type B for both halves were both significantly different (for the first half: $t = 22.38, p < 0.0001$; for the second half: $t = 28.77, p < 0.0001$): when evaluated against the *de se* + *de re* reading, participants systematically rejected the *de re* + *de re* reading throughout the experiment.

B. Discussion

The results of the split-half analysis indicate that a participant's experience may be a crucial factor influencing semantic judgments. While the experiment overall shows a dispreference for the *de re* + *de se* reading, as reported in [8], further examination of the results reveals that this dispreference is not immediately manifested. Rather, participants only gradually develop a dispreference after multiple exposures to pictures depicting such an interpretation. This finding suggests that a dispreference for the *de re* + *de se* reading may be due to factors relating to performance or experience, and not (or perhaps not only) due to stipulations provided by the grammar.

III. Study 2: Does the ORC Effect hold with other counterfactual verbs?

Following ideas first articulated in [12], it is claimed that the notion of counterfactuality determines the availability of the *de re* + *de se* reading, with recent hypotheses stating that counterfactual verbs such as *dream* prohibit it, but non-counterfactual verbs such as *believe* allow it [13], [8]. In Study 2, we test this hypothesis further by conducting an experiment with a different counterfactual verb, replacing *dream* with *imagine*, as in (3). Additionally, as with Study 1, we applied the split-half method to investigate whether the observed effects were stable across the duration of the experiment, or whether participants' semantic judgments varied as a function of experience.

(3) There were two authors, Carol and Sandra. Carol imagined that she was Sandra and **she** was buying **her** book.

A. Method

The materials, pictures, and procedure were identical to those employed in Experiment 2 in [8], except that the verb *dream* was replaced with *imagine*.

B. Results and Discussion

We used Amazon Mechanical Turk [11] for the experiments reported in Study 2 and 3. For both experiments, we only recruited participants with IP addresses based in the United States. Participants were asked to self-identify as native English speakers, and data from participants who didn't were discarded. The analyses we report below are from the data provided by the remaining participants.

Participants ($n = 104$) provided semantic judgments, which were analyzed using a binary logistic regression model with Comparison Type, presentation order, and their interaction, as predictors. Additionally, we allowed the model to be adjusted by items, subjects, and lists, in order to account for random effects. A significant effect of Comparison Type was observed ($z = -4.53, p < 0.0001$), suggesting that the pattern of responses differed depending on which pair of readings participants were presented with. Order of presentation was also a significant predictor ($z = 4.49, p < 0.0001$), indicating that the proportion of responses changed as a function of exposure, as with Study 1. Unlike Study 1, however, we also observed a significant interaction between Comparison Type and presentation order ($z = -3.00, p = 0.003$), suggesting that the effect of repeated exposure was not identical across both comparison types.

We then applied the split-half method to probe whether semantic judgments were consistent throughout the course of the experiment. We divided the dataset into two halves: the "first half" consisted of experimental items that had Presentation Orders 1 to 5 ($n = 336$), and the "second half" consisted of experimental items that had Presentation Orders 6 to 18 ($n = 348$). We analyzed each half separately using a binary logistic regression model with Comparison Type as a predictor. On both halves, there was a significant effect of Comparison Type (for the first half, $z = -6.41, p < 0.0001$; for the second half, $z = -9.34, p < 0.0001$): as with Study 1, the pattern of judgments was different in both halves when the *de re* + *de re* reading was being compared against the *de re* + *de se* reading, as opposed to against the *de se* + *de re* reading.

We then conducted pairwise T-tests within comparison type on both halves to see whether there were significant differences in the proportion of semantic judgments. Table II summarizes the proportion of semantic acceptability per condition across first and second halves, as well as for the overall experiment. For both halves, the proportion of responses in Comparison Type A was significantly different (for the first half, $t = 2.67, p = 0.01$; for the second half, $t = -4.88, p < 0.0001$): participants exhibited a preference for one reading over the other. Crucially, however, the direction of preference differed between the first and second halves: while participants significantly preferred the *de re* + *de se* reading in the first half, the opposite was true in the second half. Finally, the proportion of responses in Comparison Type B for both halves were both significantly different, and in the same direction of preference (for the first half, $t = 25.05, p < 0.0001$; for the second half, $t = 37.76, p < 0.0001$): participants systematically rejected the *de re* + *de re* reading throughout the experiment.

TABLE II. PROPORTIONS OF SEMANTIC ACCEPTABILITY ACROSS FIRST AND SECOND HALVES FOR STUDY 2

First Half	*de re+de re*	*de re+de se*	*de se+de re*	*p-value*
Comparison A	41%	58%		0.01
Comparison B	8.8%		91%	<0.0001
Second Half				
Comparison A	63.8%	36.1%		<0.0001
Comparison B	4.8%		95.1%	<0.0001
Overall				
Comparison A	51.9%	48%		0.46
Comparison B	6.7%		93.2%	<0.0001

The results of Study 2 provide further evidence of how semantic judgments can be affected by one's experiences. Using a different counterfactual verb, we have shown that participants' judgments on the acceptability of various pronoun interpretations can change as a function of continued and repeated exposure to the stimulus in question. The experiment overall shows that when forced to choose between a *de re* + *de re* and a *de re* + *de se* reading, participants are equally likely to choose either reading. However, a closer look at the pattern of results across time shows that participants initially preferred the *de re* + *de se* reading. Only after prolonged exposure did participants' judgments reverse, showing a preference for the *de re* + *de re* reading. In the context of theories that predict the unacceptability of the *de re* + *de se* reading by virtue of a grammatical stipulation, this pattern of results calls for explanation.

IV. STUDY 3: IS THE ORC EFFECT OBSERVED WITH OTHER EXPERIMENTAL PARADIGMS?

The previous studies used a two-picture forced-choice task to elicit semantic judgments. One potential confound that this paradigm raises, however, is that the semantic judgments that participants provided for a particular reading were always relative to another reading. In other words, in this paradigm, participants' choices essentially indicated which of the two readings were preferred, but preference for one reading over the other does not automatically mean that the dispreferred reading is unacceptable. In Study 3, we addressed this issue by utilizing a single-picture rating task. Instead of comparing two potential readings and indicating which reading they preferred, participants simply provided ratings of semantic acceptability on single pictures. Participants therefore are not explicitly asked to entertain two potential readings at the same time.

In addition to the change in experimental paradigm, this study also attempted to investigate the role of counterfactuality and its potential effect on the (un)availability of the *de re* + *de se* reading. As mentioned earlier, recent hypotheses state that counterfactual verbs such as *dream* and *imagine* prohibit this reading, but non-counterfactual verbs like *believe* and *claim* allow it. We experimentally address this hypothesis in the current study to investigate if there are systematic differences in semantic judgments between these two verb types.

TABLE III. MEAN RESPONSES (STANDARD DEVIATIONS IN PARENTHESES) OF SEMANTIC ACCEPTABILITY PER VERB TYPE AND TYPE OF READING IN STUDY 3

	Type of reading	
Type of verb	*de re+de se*	*de se+de re*
Counterfactual	6.03 (1.41)	5.33 (1.98)
Non-counterfactual	6.1 (1.24)	5.03 (2.21)

A. Method

The materials and pictures were identical to those employed in Study 2, except that we used 4 counterfactual verbs (*dream*, *imagine*, *pretend*, *wish*) and 4 non-counterfactual verbs (*believe*, *think*, *say*, *claim*), interspersed among 32 filler items, presented randomly. All of the verbs used in the test and filler items appeared only once in the course of the experiment, thereby preventing participants from forming verb-based response strategies due to multiple exposures. Unlike in Study 2, it was not possible to terminate the experiment until all 40 items were completed.

Participants were given discourses similar to (1) and (3) and were instructed to rate on a 7-point Likert scale how well the discourse matched the picture that was provided with it. The test items were presented either with a picture depicting the *de re* + *de se* or *de se* + *de re* reading. Half of the filler items were presented with a mismatching entity or object, so that there was expected to be a clear answer as to whether the picture fit the story or not, allowing us to determine if participants were paying attention to the task. The experiment was counterbalanced into 4 lists, such that for each test item, participants were only presented with one of the two potential readings.

B. Results and Discussion

60 participants provided semantic judgments, which were analyzed using a linear mixed-effects regression model, with counterfactuality, reading type, order of presentation, and their interactions, as fixed factors. We allowed the model to be adjusted by items, subjects, and lists, in order to account for random effects. Means and standard deviations of responses per condition are provided in Table III. The model revealed a main effect of reading type ($t = -2.35, p = 0.01$): pictures depicting the *de re* + *de se* reading were rated as more acceptable than pictures depicting the *de se* + *de re* reading. There was no main effect of counterfactuality ($t = -0.07, p = 0.94$): acceptability ratings did not differ as a function of counterfactuality. Unlike the previous two experiments, however, there was no effect of presentation order ($t = -0.29, p = 0.76$). Finally, there were no significant interactions (all ps > 0.05).

The results of Study 3 further illustrate how semantic acceptability judgments can be affected by the way these judgments are elicited. Unlike in Study 1 and 2, participants did not evaluate two potential readings and picked the more acceptable reading. Study 3 only presented participants with one reading to entertain at a time, and in this paradigm, participants exhibited a higher acceptability for the *de re* + *de se* reading than the *de se* + *de re* reading. It is also worth noting that the average scores for both readings were relatively high (all conditions were 5 or higher in a 7-point scale), suggesting that participants overall deemed both readings as acceptable. These results are inconsistent with theories predicting that the *de re* + *de se* reading is unacceptable due to grammatical stipulations.

V. GENERAL DISCUSSION

The results of the three studies reported here illustrate how semantic judgments can be influenced by one's experience. Our studies have shown that judgments of semantic acceptability can vary depending on the amount of exposure, as well as on the way these judgments are elicited. Our results pattern with recent studies on behavioral responses to syntactic and semantic stimuli, which demonstrated that a) one's syntactic parsing preferences can change in the course of a single experiment by simply manipulating the statistical distribution of frequent and non-frequent structures [14], [15]; and that b) even online behavioral responses to semantic priming paradigms can be prone to strategic effects that are allowed by the composition of the experiment [16]. The results of the studies we report highly suggest that the semantic acceptability judgments we elicited is not only affected by relatively stable knowledge of the grammar of the language, but also by factors pertaining to exposure: how often these judgments are elicited in an elicitation session, and the task employed to elicit these judgments. Our studies illustrate how one's experience affects semantic judgments: semantic judgments changed as a function of repetition, as well as of presentation.

Reviewers have drawn our attention to potential problems with the split-halves analyses we implemented in Study 1 and 2. Due to the design of the experiments, the participants in both halves are not identical (the participants in the second half form a proper subset of the participants in the first half). It was suggested instead to simply use order of presentation as a gradient variable. As reported above, we implemented this suggestion, and as expected, presentation order was a significant predictor in both Study 1 and 2. However, we believe that dividing the acceptability judgments in halves provides at least a coarse-grained index allowing us to see how semantic judgments change as a function of exposure. As Tables I and II show, the unacceptability of the *de re* + *de se* reading only emerges in the second half. Even if we discard the second halves of Study 1 and 2, the data comprising the first halves indicate that there is no dispreference for the *de re* + *de se* reading. In fact, the first half of Study 2 even shows that this reading is preferred.

Our findings have potentially serious implications for theories of pronoun interpretation that appeal to grammatical stipulations. Theories that predict the unacceptability of the *de re* + *de se* reading by virtue of a grammatical rule would be hard-pressed to explain our results, which showed that this reading was in fact not only acceptable but also preferred given certain conditions. While we do not claim that our studies provide definitive evidence against the Oneiric Reference Constraint, we believe that there is enough evidence to state that the (un)acceptability of this reading is modulated by multiple factors in addition to grammatical rules.

Our results stress how important it is to be aware of numerous factors that affect linguistic judgments. While researchers who work solely in a theoretical domain may draw

a line between linguistic knowledge/competence and linguistic performance, linguistic behavior may be affected by both linguistic and non-linguistic factors – a possibility that must be taken into consideration in experimentally-oriented research [17], [18]. This point carries particular significance when the goal is to formulate theoretical hypotheses on the basis of experimental evidence. Our findings illustrate the importance of converging evidence from multiple tasks and paradigms when testing and formulating theoretical hypotheses.

ACKNOWLEDGMENT

The authors would like to thank Martin Herfurth for assistance with programming and data collection. The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7 2007-2013) under grant agreement No. 618871.

REFERENCES

[1] L. A. Jakobovits, "Effects of repeated presentation on cognitive aspects of behavior: Some experiments on the phenomenon of semantic satiation," Ph.D. dissertation, McGill University, 1962.

[2] ——, "Semantic satiation and cognitive dynamics," *Journal of Special Education*, vol. 2, pp. 35–44, 1967.

[3] W. Snyder, "An experimental investigation of syntactic satiation," *Linguistic Inquiry*, vol. 31, no. 3, pp. 575–582, 2000.

[4] K. Hiramatsu, "Accessing linguistic competence: Evidence from children's and adult's acceptability judgments," Ph.D. dissertation, University of Connecticut, 2000.

[5] R. P. Chaves and J. E. Dery, "Which subject islands will the acceptability of improve with repeated exposure?" in *Proceedings of the 31st West Coast Conference on Formal Linguistics*, R. E. Santana-LaBarge, Ed., Somerville, MA, 2014, pp. 38–45.

[6] J. Sprouse, "Revisiting satiation: Evidence for an equalization response strategy," *Linguistic Inquiry*, vol. 40, no. 2, pp. 329–341, 2009.

[7] O. Percus and U. Sauerland, "Pronoun movement in dream reports," in *NELS 33: Proceedings of the 33rd Annual Meeting of the North East Linguistics Society*, S. Kawahara and M. Kadowaki, Eds., Cambridge, MA, 2003, pp. 347–366.

[8] H. Pearson and J. E. Dery, "Dreaming *de re* and *de se*: Experimental evidence for the Oneiric Reference Constraint," in *Proceedings of Sinn und Bedeutung 18*, U. Etxeberria, A. Fălăuş, A. Irurtzun, and B. Leferman, Eds., Vitoria-Gasteiz, 2013, pp. 322–339.

[9] R. A. Harshman and W. S. De Sarbo, "An application of PARAFAC to a small sample problem, demonstrating preprocessing, orthogonality constraints, and split-half diagnostic techniques," in *Research methods for multimode data analysis*, H. G. Law, C. W. Snyder, J. A. Hattie, and R. P. McDonald, Eds. New York, NY: Praeger, 1984, pp. 602–642.

[10] K. R. Murphy and C. O. Davidshofer, *Psychological testing: Principles and applications*. Upper Saddle River, NJ: Pearson/Prentice Hall, 2005.

[11] W. Mason and S. Suri, "Conducting behavioral research on Amazon's Mechanical Turk," *Behavior Research Methods*, vol. 44, no. 1, pp. 1–23, 2012.

[12] O. Percus, "Uninterpreted pronouns?" unpublished.

[13] H. Pearson, "The sense of self: Topics in the semantics of *de se* expressions," Ph.D. dissertation, Harvard University, 2013.

[14] A. B. Fine, T. F. Jaeger, T. A. Farmer, and T. Qian, "Rapid expectation adaptation during syntactic comprehension," *PLoS ONE*, vol. 8, no. 10, pp. 1–18, 2013.

[15] T. A. Farmer, A. B. Fine, S. Yan, S. Cheimariou, and T. F. Jaeger, "Error-driven adaptation of higher-level expectations during reading," in *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, Quebec City, 2014.

[16] C. E. Su, "When is semantic priming automatic? Instrument and location participant role priming as a case study," Ph.D. dissertation, State University of New York at Buffalo, 2012.

[17] J. Sprouse, S. Fukuda, H. Ono, and R. Kluender, "Reverse island effects and the backward search for a licensor in multiple *wh*-questions," *Syntax*, vol. 14, no. 2, pp. 179–203, 2011.

[18] J. Sprouse, I. Caponigro, C. Greco, and C. Cecchetto, "Experimental syntax and the variation of island effects in English and Italian," *Natural Language and Linguistic Theory*, 2015.