Grounded and Ungrounded Referring Expressions in Human Dialogues: Language Mirrors Different Grounding Conditions

Eleonora Gualdoni, Raffaella Bernardi
University of Trento
eleonora.gualdoni@studenti.unitn.it
raffaella.bernardi@unitn.it

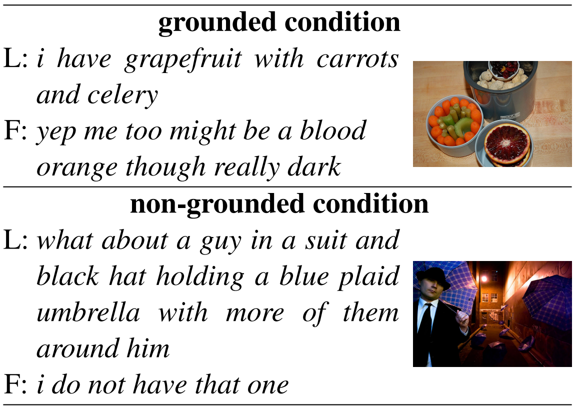
Raquel Fernández, Sandro Pezzelle
University of Amsterdam
raquel.fernandez@uva.nl
s.pezzelle@uva.nl

Abstract

We study how language use differs between dialogue partners in a visually grounded reference task when a referent is mutually identifiable by both interlocutors vs. when it is only available to one of them. In the latter case, the addressee needs to *disconfirm* a proposed description – a skill largely neglected by both the theoretical and the computational linguistics communities. We consider a number of linguistic features that we expect to vary across conditions. We then analyze their effectiveness in distinguishing among the two conditions by means of statistical tests and a feature-based classifier. Overall, we show that language mirrors different grounding conditions, paving the way to future deeper investigation of referential disconfirmation.

1 Introduction

Communication is a joint activity in which interlocutors share or synchronize aspects of their private mental states and act together in the world. To understand what our minds indeed do during communication, Brennan et al. (2010) highlight the need to study language in interpersonal coordination scenarios. When a conversation focuses on objects, interlocutors have to reach the mutual belief that the addressee has identified the discussed referent by means of visual grounding. In this frame, Clark and Wilkes-Gibbs (1986) have pointed to *referring* as a collaborative process, that requires action and coordination by both speakers and interlocutors, and that needs to be studied with a collaborative model. Clark and Wilkes-Gibbs (1986), in fact, have highlighted that – in order to refer to an object in the world – speakers must believe that the referent is *mutually identifiable* to them and their addressees. This is an important skill that human speakers leverage to succeed in communication.



grounded condition

L: *i have grapefruit with carrots and celery*
F: *yep me too might be a blood orange though really dark*

non-grounded condition

L: *what about a guy in a suit and black hat holding a blue plaid umbrella with more of them around him*
F: *i do not have that one*

Figure 1: Examples of dialogue segments where the image referent is visible to both leader and follower (grounded condition) or only visible to the leader (non-grounded condition).

However, humans are not only able to identify an object described by the interlocutor – that is, *grounding* a referring expression – but also to *understand that such an object is not in the scene and, therefore, it cannot be grounded.* It can happen, indeed, that a referent is not mutually identifiable by the speakers, due to the speakers being in different grounding conditions. In this case, the addressee is able to *disconfirm a description* stated by the interlocutor by communicating that he/she does not see it (as in Figure 1). This is a crucial skill of human speakers. However, it is often neglected in the computational modelling of conversational agents.

We conjecture that the participants' visual grounding conditions have an impact on the linguistic form and structure of their utterances. If confirmed, our hypothesis would lead to the claim that mature AI dialogue systems should learn to master their language with the flexibility shown by humans. In particular, their language use should differ when the referred object is mutually identifiable or not. It has been shown that current AI multimodal systems are not able to decide if a visual question is answerable or not (Bhattacharya et al., 2019), and they fail to identify whether the entity to which an expression refer is present in the visual scene or not (Shekhar et al., 2017b; Shekhar et al., 2017a). We believe models can acquire this skill if they learn to play the "language game" properly.

In this paper, we investigate how the language of human conversational partners changes when they are in a mutually grounded (they both see the image they are speaking about) or non-mutually grounded setting (one sees the image while the other does not).

We find that, indeed, there are statistically significant differences along various linguistic dimensions, including utterance length, parts of speech, and the degree of concreteness of the words used. Moreover, a simple SVM classifier based on these same features is shown to be able to distinguish between the two conditions with a relatively high performance.

2 Dataset

We take the PhotoBook dataset (Haber et al., 2019) as our testbed: two participants play a game where each sees a different grid with six images showing everyday scenes.[1] Some of the images are common to both players, while others are only displayed to one of them. In each grid, three of the images are highlighted. By chatting with their dialogue partner, each player needs to decide whether each of the three highlighted images is also visible to their partner or not.

A full game consists of five rounds, and the players can decide to move to the next round when they are confident about their decisions. As the game progresses, some images may reappear in subsequent rounds. The corpus is divided into dialogue *segments*: the consecutive utterances that, as a whole, discuss a given target image and include expressions referring to it. From the set of all segments in PhotoBook, we create our dataset by focusing on segments belonging to the first round of a game (since at that point all images are new to the participants) and where a single image is being discussed.[2] This results in a dataset composed of 3,777 segments paired with a given image referent and an action label indicating whether the referent is visible to both participants or only to one. The annotated dataset, together with other relevant materials, is available at: `https://dmg-photobook.github.io/`

The PhotoBook task does not impose a specific role on the players, unlike for example the MapTask corpus (Anderson et al., 1991), where there are predefined *information giver* and *information follower* roles. In PhotoBook, the dialogues typically follow this scheme: one of the participants spontaneously decides to describe one of the images highlighted in their grid and the other participant indicates whether they also have it in their own grid or not. We call the former player the *leader* and the latter the *follower*.[3] We refer to situations where the follower also sees the image described by the leader as the **grounded condition** and those where the follower does not see the image as the **non-grounded condition**. Naturally, the leader always sees the referent image.

Out of the 3,777 dialogue segments in our dataset, 1,624 belong to the grounded condition and 2,153 to the non-grounded one.

3 Linguistics Features

We hypothesize that the language used by the dialogue participants will differ in the grounded vs. non-grounded condition. To test this hypothesis, we first identify several linguistic features that we expect to vary across conditions.

Length. We expect that the length of the utterances and the overall dialogue segments may depend on the players' possibility to see the referent. For example, in the non-grounded condition more utterances may be needed to conclude that the follower does not see the referent (thus leading to longer segments). Furthermore, not seeing the referred image could limit the expressivity of the utterances by non-grounded follower (thus leading to shorter utterances).

[1] The images used in the PhotoBook task are taken from the MS COCO 2014 Trainset (Lin et al., 2014).

[2] We discard segments that refer to more than one image as well as those labelled with the wrong image by the original heuristics (Haber et al., 2019).

[3] We use simple heuristics to assign these roles a posteriori: when the image is not in common, we label as the follower the participant who does not see the image, while when the image is visible to both participants we consider the follower the player who produces the last utterance of the segment. We manually corrected the classification of the few segments that did not follow this general rule.

We compute utterance length as number of tokens per utterance and segment length as both number of tokens per segment and number of utterances per segment.

Word frequency. Frequency effects are key in psycholinguistics. Word frequency is one of the strongest predictors of processing efficiency (Monsell et al., 1989) and experiments have confirmed its link to memory performances (Yonelinas, 2002). It is plausible that different grounding conditions lead to different word choices, and that word frequency turns out to be a key aspect of this linguistic variation.

To estimate word frequency, we use off-the-shelf lemma frequency scores (frequency per million tokens) from the British National Corpus (Leech et al., 2014).[4] For each segment in our dataset, we compute the average word frequency by first lemmatizing the words in the segment and then calculating the average frequency score for all lemma types in the segment.[5]

Concreteness. Concreteness is fundamental to human language processing since it helps to clearly convey information about the world (Hill and Korhonen, 2014). We use the *concreteness scores* by Brysbaert et al. (2014), corresponding to 40K English word lemmas, and collected via crowd-sourcing, where participants were requested to evaluate *word-concreteness* by using a 5-point rating scale ranging from abstract to concrete. We compute the average word concreteness by first lemmatizing the words in the segment and then calculating the average score for all lemma types in the segment without repetitions, divided by part-of-speech (POS).[6]

Parts of Speech distributions. Different POS differ in their function and descriptive power. We thus expect that their distribution will vary between grounded and non-grounded conditions. For example, we expect *nouns* and *adjectives* to be more likely in visually grounded referential acts, while determiners may signal whether the referent is in common ground or not (*the* vs. *a*) and give clues about the polarity of the context where they are used (*any* vs. *each*).

We extract POS distributions by first POS-tagging the utterances in the dataset[7] and then computing the proportion of words per segment that are nouns, adjectives, verbs, or determiners, respectively. Given the different functions of different determiners, we break down this class and independently compute proportions for each of the following determiners: *a/an, the, that, those, this, these, some, all, each, any, half, both.*

4 Statistical Analysis

To test our hypothesis that the language used by the participants differs in the grounded vs. non-grounded condition, we perform a statistical analysis on our data. We compare: (1) the utterances by the *leaders* in the grounded and non-grounded conditions, and (2) the utterances by the *followers* in the grounded and non-grounded conditions. We evaluate the statistical significance of these comparisons with a Mann-Whitney U Test, which does not assume the data fits any specific distribution type. Below we report the results of each of these comparisons. Unless otherwise specified, statistical significance is tested for $p < 0.001$.

Length. Followers use significantly fewer words while leaders use significantly more words in the non-grounded condition than in the grounded condition. This trend is also illustrated in the example in Figure 1. Although followers use fewer words in the non-grounded condition, they produce a significantly higher number of utterances per segment, while no reliable differences are observed for the leaders (see Figure 2a and 2e, respectively). These findings indicate that establishing that a referring expression cannot be commonly grounded requires more evidence and more information than resolving the expression.

Frequency. Followers use significantly more high-frequency words in the grounded condition than the non-grounded condition, in particular for nouns and conjunctions. This is consistent with the reported production of more utterances per segment in the non-grounded condition, and suggests that the non-grounded follower uses them to talk about fine-grained details described by low-frequency words. In contrast, high-frequency verbs are reliably more common in the non-grounded condition (see Figure 2b).

[4] Available at `http://ucrel.lancs.ac.uk/bncfreq/flists.html`

[5] Lemmas not present in the BNC lists are ignored.

[6] Lemmas not present in the corpus are ignored.

[7] We use the NLTK Python library (Bird et al., 2009) in its "universal" tagset version.

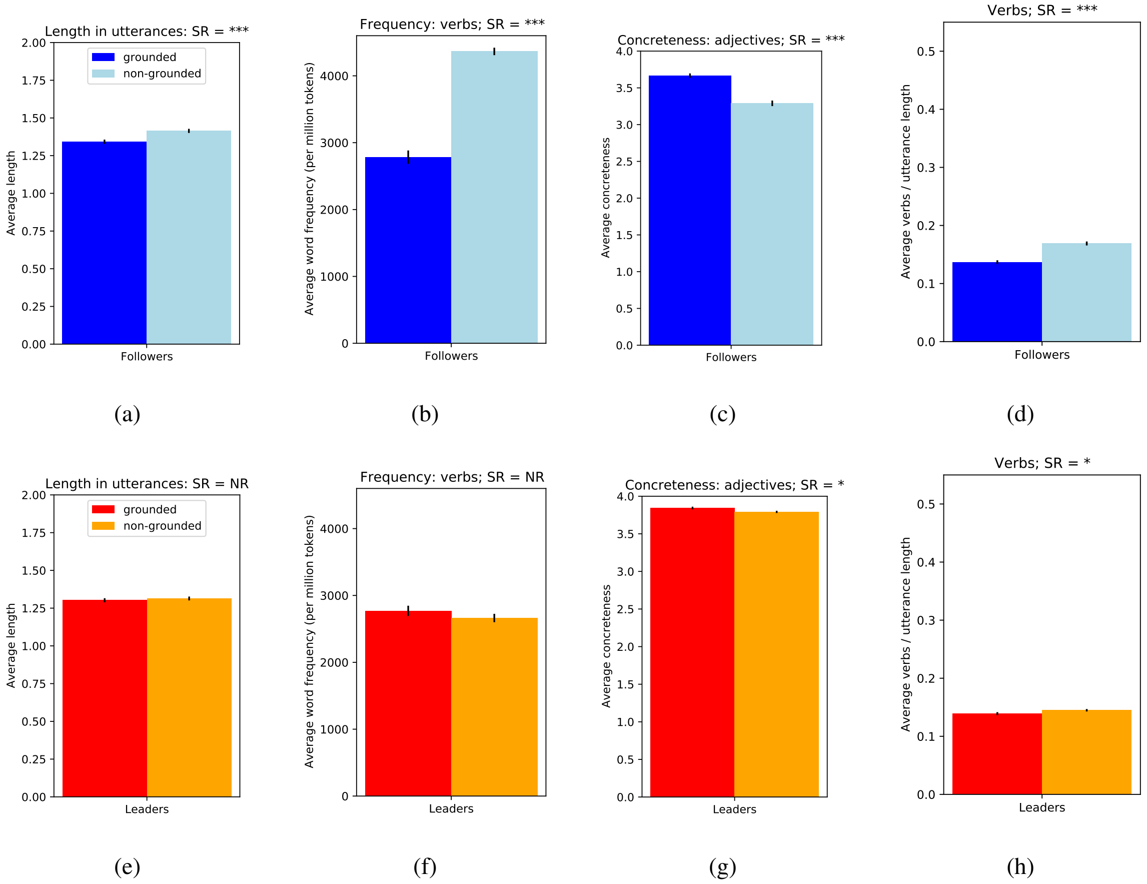


Figure 2: From left to right, difference between grounded and non-grounded condition for: (a/e) number of utterances per segment; (b/f) frequency of used verbs; (c/g) concreteness of used adjectives; (d/h) proportion of verbs. Top: followers; bottom: leaders. We use *** to refer to statistical significance at $p < 0.001$; ** for $p < 0.01$; * for $p < 0.05$; . for $p < 0.1$. Best viewed in color.

For example, note the high-frequency verbs *do* and *have* used by the non-grounded follower in Figure 1. The language of leaders, in contrast, shows marginally reliable or no difference across conditions regarding word frequency (see, e.g., the case of verbs in Figure 2f), except for high-frequency nouns and conjunctions, which are reliably more common in the grounded condition ($p < 0.01$).

Concreteness. Somehow counterintuitively, followers use overall significantly more concrete words in the non-grounded than in the grounded condition. However, an opposite pattern is found for adjectives, which usually describe the colors of the objects in the scene (see Figure 2c). This latter result is in line with our intuitions: in the non-grounded condition, followers do not have direct access to the specific perceptual properties of the entities in the image and hence use less concrete adjectives. As for the leaders, while nouns are reliably different, for the other POS there is either no or marginally reliable difference (see adjectives in Figure 2g, adverbs, conjunctions, and numerals) between the two conditions. This is expected since their language is always visually grounded.

Parts of speech. Followers use significantly more nouns and the determiners *a/an, the, each* in the grounded condition, while in the non-grounded condition they use significantly more verbs (see Figure 2d) and determiners *all* and *any*. That is, the grounded condition leads followers to more directly describe what they see by focusing on a specific object, as in the grounded example in Figure 1. In contrast, the non-grounded condition elicits utterances with more 'confirmation' verbs such as *do* and *have* and a more *vague* language signalled by the use of quantifiers, e.g., *"I don't have any of a cake"*. As for the leaders, we observe a mixed pattern of results, though, overall, there are less reliable differences between the two conditions compared to the followers (see the case of verbs in Figure 2h).

5 Automatic Classification

To more formally investigate the effectiveness of our selected features in distinguishing between various grounding conditions, we feed them into an SVM classifier which predicts GFC or NGFC. We run two SVM models: one for leaders, *SVM leaders*, and one for followers, *SVM followers*.[8] Our hypothesis is that *SVM leaders* should not be very effective in the binary classification task since the language of the leaders differs only on few aspects, and less reliably between the two conditions compared to the followers'. In contrast, we expect *SVM followers* to achieve a good performance in the task, given the significant differences observed between the two conditions.

Starting from all our linguistic features (see above), we excluded those that turned out to be multicollinear in a Variance Inflation Factor test (VIF).[9] The resulting N features (27 for the leaders, 28 for the followers), were used to build, for each datapoint, an N-dimensional vector of features that was fed into the classifier. We performed 10-fold cross-validation on the entire dataset.

Table 1 reports the accuracy, precision, recall and F1-score of the two SVM models. While *SVM leaders* is at chance level, *SVM followers* achieves a fairly high performance in the binary classification task. This indicates that our linguistic features are effective in distinguishing among the two conditions in the followers' segments. These results confirm that the language of the speakers in the *follower* role is affected by their grounding condition, and that a well-informed model is able to capture that by means of their language's linguistic features.

Table 2 reports the confusion matrices produced by our SVM models after 10-fold cross-validation. We can notice that *SVM leaders* wrongly labels NGFC datapoints as GFC in 1,381 cases, thus producing a high number of false positives. This does not happen with *SVM followers*, which is overall more accurate.

6 Related Work

Current multimodal systems are trained to process and relate modalities capturing correspondences between "sensory" information (Baltrusaitis et al., 2017). It has been shown they have trouble deciding if a question is *answerable* or not (Bhattacharya et al., 2019). Moreover, they fail to identify whether the entity to which an expression refers is present in the visual scene or not (Shekhar et al., 2017b; Shekhar et al., 2017a). Connected to this weakness is the limitation they encounter when put to work as dialogue systems, where they fail to build *common ground* from minimally-shared information (Udagawa and Aizawa, 2019). To be successful in communication, speakers are supposed to attribute mental states to their interlocutors even when they are different from their own (Rabinowitz et al., 2018; Chandrasekaran et al., 2017). This, in multimodal situations, can happen when the visual scene is only partially common between them. AI models have difficulties in such conditions (Udagawa and Aizawa, 2019).

We study how the language of conversational partners changes when (i) speakers refer to an image their interlocutor does not see and (ii) neither of the two is aware of this unshared visual ground. Though the idea that the grounding conditions of the addressees can affect their interlocutor's language is not new in psycholinguistics (Brennan et al., 2010; Brown and Dell, 1987; Lockridge and Brennan, 2002; Bard and Aylett, 2000), our approach differs from previous ones since it proposes a computational analysis of visual dialogues. Moreover, differently from other computational approaches (Bhattacharya et al., 2019; Gurari et al., 2018), we investigate scenarios where the disconfirmation of a referent's presence *is the answer* instead of suggesting a case of *unanswerability*.

7 Conclusion

Our findings confirm that, in a visually-grounded dialogue, different linguistic strategies are employed by speakers based on different grounding conditions. Our statistical analyses reliably indicate that several aspects of the language used in the conversation mirror whether the referred image is – or not – *mutually shared* by the interlocutors. Moreover, the effectiveness of a simple feature-based classifier to distinguish between the two followers' conditions further indicates that the lan-

[8] We experiment with the `scikit-learn` Python library (Pedregosa et al., 2011) for C-Support Vector Classification. We use the default Radial Basis Function (`rbf`) kernel. Parameter C set to 100 gives the best results.

[9] The VIF test indicates whether there is a strong linear association between a predictor and the others (Pituch and Stevens, 2016). When the VIF index exceeded 10, we performed a variable deletion (Myers, 1990).

	Accuracy	**Precision**			**Recall**			**F1-score**		
		GFC	**NGFC**	**Av.**	**GFC**	**NGFC**	**Av.**	**GFC**	**NGFC**	**Av.**
SVM leaders	0.57	0.15	0.89	0.40	0.50	0.58	0.55	0.23	0.70	0.50
SVM followers	**0.80**	0.77	0.79	**0.78**	0.73	0.82	0.78	0.75	0.80	**0.78**

Table 1: *Accuracy*, *Precision*, *Recall*, and *F1-score* of our SVM models, computed per class on a 10-fold cross-validation, with the corresponding weighted averages (Av.). Since our two classes (GFC and NGFC) are not balanced, chance level is 0.57.

	SVM leaders		**SVM followers**	
	GFC	**NGFC**	**GFC**	**NGFC**
GFC	243	1381	1245	379
NGFC	242	1911	461	1692

Table 2: The confusion matrices produced by our SVM models on a 10-fold cross-validation.

guage used by the speakers differs along several dimensions. We believe this capability of humans to flexibly tune their language underpins their success in communication. We suggest that efforts should be put in developing conversational AI systems that are capable to master language with a similar flexibility. This could be achieved, for example, by exposing models to one or the other condition during training to encourage them encode the relevant linguistic features. Alternatively, they should first *understand* whether the grounded information which is referred to is available to them or not. These are open challenges that we plan to tackle in future work.

Acknowledgments

EG carried out part of the work while being an ERASMUS+ visiting student at the University of Amsterdam. SP and RF are funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455 awarded to RF).

References

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC map task corpus. *Language and speech*, 34(4):351–366.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. Multimodal machine learning: A survey and taxonomy. *CoRR*, abs/1705.09406.

Ellen G Bard and MP Aylett. 2000. Accessibility, duration, and modeling the listener in spoken dialogue. In *Proceedings of the Götalog 2000 Fourth Workshop on the Semantics and Pragmatics of Dialogue*.

Nilavra Bhattacharya, Qing Li, and Danna Gurari. 2019. Why does a visual question have different answers? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4271–4280.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Susan E Brennan, Alexia Galati, and Anna K Kuhlen. 2010. Two minds, one dialog: Coordinating speaking and understanding. In *Psychology of learning and motivation*, volume 53, pages 301–344. Elsevier.

Paula M Brown and Gary S Dell. 1987. Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology*, 19(4):441 – 472.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46(3):904–911.

Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. 2017. It takes two to tango: Towards theory of AI's mind. *CoRR*, abs/1704.00717.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández.

2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910.

Felix Hill and Anna Korhonen. 2014. Concreteness and subjectivity as dimensions of lexical meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 725–731.

Geoffrey Leech, Paul Rayson, et al. 2014. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.

Calion Lockridge and Susan Brennan. 2002. Addressees' needs influence speakers' early syntactic choices. *Psychonomic bulletin & review*, 9:550–7, 10.

Stephen Monsell, Michael C Doyle, and Patrick N Haggard. 1989. Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, 118(1):43.

Raymond H Myers. 1990. *Classical and modern regression with applications*. Duxbury, Boston, MA, 2nd edition.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Keenan A Pituch and James P Stevens. 2016. *Applied Multivariate Statistics for the Social Sciences*. Routledge, 6th edition.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *International Conference on Machine Learning*, pages 4218–4227.

Ravi Shekhar, Sandro Pezzelle, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017a. Vision and language integration: Moving beyond objects. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017b. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265.

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. *CoRR*, abs/1907.03399.

Andrew P Yonelinas. 2002. The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, 46(3):441–517.