Language Resources and Evaluation (2023) 57:1081–1105
https://doi.org/10.1007/s10579-022-09619-y

ORIGINAL PAPER

Spontaneous, controlled acts of reference between friends and strangers

Sean Trott[1] · **Benjamin Bergen**[1] · **Eva Wittenberg**[2]

Accepted: 15 September 2022 / Published online: 28 November 2022

Abstract

Speakers enjoy considerable flexibility in how they refer to a given referent—referring expressions can vary in their form (e.g., "she" vs. "the cat"), their length (e.g., "the (big) (orange) cat"), and more. What factors drive a speaker's decisions about how to refer, and how do these decisions shape a comprehender's ability to resolve the intended referent? Answering either question presents a methodological challenge; researchers must strike a balance between experimental control and ecological validity. In this paper, we introduce the SCARFS (Spontaneous, Controlled Acts of Reference between Friends and Strangers) Database: a corpus of approximately 20,000 English nominal referring expressions (NREs), produced in the context of a communication game. For each NRE, the corpus lists the concept the speaker was trying to convey (from a set of 471 possible target concepts), formal properties of the NRE (e.g., its length), the relationship between the interlocutors (i.e., friend vs. stranger), and the communicative outcome (i.e., whether the expression was successfully resolved). Researchers from diverse disciplines may use this resource to answer questions about how speakers refer and how comprehenders resolve their intended referent—as well as other fundamental questions about dialogic speech, such as whether and how speakers tailor their utterances to the identity of their interlocutor, how second-degree associations are generated, and the predictors of communicative success.

Keywords Reference · Common ground · Conversation · Natural Language Processing

✉ Sean Trott
sttrott@ucsd.edu

[1] Department of Cognitive Science, UC San Diego, San Diego, United States

[2] Department of Linguistics, UC San Diego, San Diego, United States

1 Introduction

Nearly all linguistic utterances involve reference. Often, reference takes the form of **nominal referring expressions (NREs)**, a category including Noun Phrases (e.g., "the cat") and pronouns (e.g., "she"). Speakers enjoy considerable flexibility in how much detail they can provide about a given referent. NREs can vary in their form (e.g., full noun phrases vs. pronouns), their length and level of modification (e.g., "the (big) (orange) cat"), their level of specificity (e.g., "the tabby" vs. "the mammal"), and more. Despite much research, it is still unclear which factors drive a speaker's decisions about how to refer, and how these decisions influence a comprehender's ability to successfully resolve that reference.

Both questions have received considerable attention, not only from theoreticians interested in systematizing a formal model of reference (Gundel et al., 1993; Tily & Piantadosi, 2009; Clark & Brennan, 1991; Schegloff & Sacks, 1979), but also among Natural Language Processing (NLP) practitioners hoping to develop improved models of reference resolution (Sukthanker et al., 2020; Zheng et al., 2011) and referring expression generation (Dale & Reiter, 1995; Gatt et al., 2007; Kunze et al., 2017; Viethen & Dale, 2006; Williams & Scheutz, 2017). Successful models of reference resolution and generation could also be critical for applications such as empathetic dialogue systems (Ma et al., 2020).

One key issue within this field of inquiry concerns the role of *common ground*, the set of mutual beliefs and shared knowledge across interlocutors (Clark & Brennan, 1991), in shaping both the production and comprehension of referring expressions. If a speaker knows that their interlocutor is already familiar with the concept at hand, less elaboration or explanation may be required, allowing for the use of shorter, more efficient NREs. Indeed, a number of findings support the hypothesis that common ground plays an important role in the production and comprehension of NREs (Brennan & Clark, 1996; Winters et al., 2018; Hawkins et al., 2020; Isaacs & Clark, 1987; Brown-Schmidt, 2012); however, other work (Horton & Keysar, 1996; Keysar et al., 1998; Fussell & Krause, 1989; Pollmann & Krahmer, 2018; Schober & Carstensen, 2010) has produced more mixed results, revealing important constraints on the magnitude and timing of this influence.

This question has practical consequences as well: if common ground *does* facilitate efficient and effective reference, then NLP systems might benefit from constructing explicit models of their interlocutors' beliefs and knowledge states—as some have already argued (Ma et al., 2020). If, on the other hand, real-time conversation is guided more by local cues and heuristics (Shintel & Keysar, 2009), then NLP systems might benefit from simply modeling those heuristics.

However, answering any of these questions faces a major methodological challenge: striking a balance between experimental control and ecological validity. Psycholinguistic experiments (Horton & Keysar, 1996; Hanna et al., 2003; Keysar et al., 2000) afford greater control over both referents and NREs, allowing researchers to investigate the causal mechanisms underlying NRE processing and production, but it is not always clear whether these results generalize to reference "in the wild". Studies of naturally-occurring conversations have greater ecological validity (Schegloff & Sacks, 1979; Schegloff, 1996), but due to their observational nature, they do not nec-

essarily allow researchers to make causal inferences. Finally, computational models (Williams & Scheutz, 2017; Dale & Reiter, 1995) provide a useful testbed for theory, but typically require some form of evaluation criteria, which raises the question of whether their performance ought to be compared to human behavior elicited in the lab or observed in more naturalistic settings.

In the current work, we introduce and describe the **SCARFS (Spontaneous, Controlled Acts of Reference between Friends and Strangers) Database**, which attempts to strike this balance between experimental control and naturalistic behavior. The corpus contains over 19,000 NREs produced in an interactive communication game, based on the word guessing game "Taboo". Unlike in fully open-ended dialogue corpora, the goal of the game is to successfully resolve reference to a predetermined word, which is listed in the corpus next to each NRE.

The corpus contains reference to 471 possible words. Each NRE produced to describe these concepts is also associated with a variable indicating whether the concept was ultimately communicated successfully, or whether the listener failed to resolve the intended meaning. Finally, as part of the task, each speaker produced NREs for two different listeners (a friend and a stranger), allowing researchers to both control for within-speaker effects and ask whether the identity of the listener—and thus, the degree of Common Ground in their interaction— impacts the NREs produced. Beyond questions about reference, this manipulation can also enable researchers to investigate other questions about how speech generally differs as a function of a speaker's relationship to their interlocutor (e.g., whether disfluencies distribute differently), how speech changes over repeated interactions with a particular interlocutor, and how second-degree associations are created.

In Sect. 2 below, we briefly review related datasets. In Sect. 3, we describe the creation of the dataset, descriptive statistics about the corpus, and the outcome of several pre-registered analyses investigating the role of common ground in reference. Finally, in the General Discussion, we explore possible use cases for the corpus.

2 Related datasets

In the past 15 years, a number of corpora have been developed to study the use of referring expressions and other properties of language use. These corpora range from annotated news articles, such as OntoNotes (Weischedel et al., 2008), to open-ended dialogues, such as CallHome American English (Canavan et al., 1997). Here, we focus primarily on English dialogue corpora, which are the most numerous.

2.1 Open-ended dialogue Corpora

Open-ended dialogues between two or more parties can be used to investigate properties of referring expressions in more naturalistic settings. Many of these involve spoken communication over the telephone—in English alone, corpora of telephone conversations include CallHome (Canavan et al., 1997), CallFriend (Canavan et al., 1996), and Switchboard (Godfrey et al., 1992). The former two corpora contain transcriptions of unscripted dialogues between a participant and a family member or

friend about a topic of their choosing; both have also been extended to other languages, such as Japanese (Wheatley et al., 1996). Switchboard contains transcriptions of telephone conversations between strangers about an assigned topic; it has also been annotated for the syntactic constructions that appear in a given utterance, as well as the conversational function of that utterance, i.e., the dialogue act (Stolcke et al., 2000). Corpora like Switchboard have been used for a number of research questions, including designing systems for natural language understanding and generation (Hu et al., 2018).

Another example of open-ended dialogue corpora is CHILDES (MacWhinney, 2000), or Child Language Data Exchange System: a collection of transcribed interactions between adults and children. CHILDES is not limited to varieties of English alone; other languages include French, Chinese, Japanese, Spanish, and more. CHILDES has been used to answer a diverse set of questions relating to language acquisition and child-directed speech. Most relevantly, it has been used to study the factors underlying a speaker's choice of referring expression (Orita et al., 2015) and the acquisition and use of referring expressions by children (Gundel et al., 2007).

Finally, the ICSI Meeting Corpus (Janin et al., 2003) contains transcriptions of recorded meetings at the International Computer Science Institute over a three-year timespan. The corpus includes speech data, word-level transcriptions, and metadata about the participants involved. It has also been further annotated for references to meeting participants (Niekrasz & Moore, 2010); these annotations can be used to aid language understanding systems with tasks such as reference resolution.

2.2 Task-Oriented dialogues

Corpora involving dialogic interactions centered around a **collaborative task** afford greater experimental control than the more open-ended corpora described above, though this sometimes comes at the cost of constraining the space of possible utterances and topics under discussion. Some of these corpora involve keyboard-mediated communication. For example, the COCONUT corpus (Di Eugenio et al., 2000) contains written interactions between pairs of participants tasked with designing the living and dining rooms of a house; among other things, it is annotated for dialogue features such as the intended function of a particular utterance. More recently, Hawkins et al. (2020) released a corpus of over 15,000 dyadic interactions, in which pairs of participants attempted to coordinate on a given referent; because this corpus involves multiple interactions between each pair, it has been used to study the role of locally constructed common ground in facilitating efficient, context-sensitive reference.

Other task-oriented corpora involve reference in a multimodal setting. The Joint Construction Task corpus (Foster et al., 2008) contains transcriptions of participants collaborating on a puzzle, as well as measures of each participant's eye-gaze during the task. The REX corpora (Takenobu et al., 2012) rely on a similar communication game as Hawkins et al. (2020), but feature pairs of participants collaborating side-by-side in the lab, separated by a dividing screen. Speech data was transcribed, annotated for various referring expression attributes, and aligned with extra-linguistic information such as eye gaze. More recently, Cirik et al. (2020) asked participants to describe

navigation to referents in panoramic images; these descriptions were used to create the Refer360 dataset, which is well-suited to studying the generation and resolution of referring expressions in the context of spatial navigation.

3 Corpus description

Our primary goal was to develop a corpus that struck a balance between the open-ended and experimentally controlled approaches reviewed above. To that end, we adapted the word guessing game Taboo for a laboratory study that elicited spontaneous speech about target concepts between friends and strangers. Taboo has been used in several previous psycholinguistic experiments to study mentalizing in the context of communication (de Boer et al., 2013; Willems et al., 2010), the role of gesture in communicating abstract concepts (Zdrazilova et al., 2018), and the effects of common ground on communicative efficacy (Pollmann & Krahmer, 2018). Like Pollmann and Krahmer (2018), we manipulated pre-existing and locally constructed common ground. In this way, we created a data set that allows for the investigation of how common ground affects the degree of efficiency and effectiveness in interlocutors' communication. We tightly controlled for the effects of common ground by making it a within-subject factor: each participant played the game with both a friend and a stranger. As demonstrated below, the dataset can also be used to answer questions about how features of the target concept itself (e.g., its concreteness or frequency) influence the probability of successful resolution.

Below, we describe the experimental procedure, the pipeline for processing and annotating each session, descriptive statistics characterizing the dataset, and the results of initial statistical analyses. The corpus itself is listed on OSF (https://osf.io/pxqvb/) and GitHub (https://github.com/seantrott/scarfs).

3.1 Experiment sescription

Participants. Individual participants were recruited through the UC San Diego Psychology Undergraduate Subject Pool. Half of participants were compensated with course credit; each of these was also asked to bring a friend, who received $18 for participating. Each session required two such pairs of friends (four participants total), so we scheduled two pairs of participants per time slot.

We ran 19 sessions altogether before shelter-in-place measures due to the COVID-19 pandemic were put in place. As per our pre-registration (https://osf.io/hcya6), we excluded two sessions from the final dataset and accompanying analyses because of an error with the audio recording. Thus, the final dataset includes 17 sessions for a total of 68 participants. Of these 68, 51 self-identified as female (16 male, 1 non-binary). (After removing three non-native speakers for the analyses described below, there was a total of 65 participants; the full dataset includes the non-native speakers, along with the necessary information for excluding or including them.)

Materials. To create all items, we modified a standard Taboo game for our purposes. As target words, we included only common nouns referring to entities, events,

or attributes. There were 471 cards altogether, distributed among 4 "decks" that we refer to by their colors: blue (117), purple (118), red (118), and yellow (118).

As in the original game, each target word was accompanied by five "taboo" words, which the speaker was not allowed to use when describing the target word. These taboo words were thematically or taxonomically related to the target word; for example, the taboo words for the card "luggage" were *baggage*, *suitcase*, *travel*, *vacation*, and *pack*. Taboo words were adapted from the original game.

Procedure. Once all four participants arrived, they were seated at the same table and paired with either their friend or a randomly selected stranger from the opposite pair; sessions were counterbalanced according to whether participants played first with their friend or a stranger. Two experimenters then explained the rules of the game. On each turn, the Speaker's job was to accurately communicate as many cards from their deck as possible, without using any of the taboo words associated with each target word. Here, "success" counted as the speaker's partner (the Guesser) correctly guessing the target word within 30 s. Each turn was 2 min. (Note that we imposed the 30 s limit on each card to ensure that Speakers/Guessers communicated about at least four cards per turn. Once 30 s had passed, Speakers/Guessers were required to communicate about a new card, and the previous card was counted as "Out of Time".)

Speakers were each assigned to a Deck (Blue, Purple, Red, Yellow), and the cards they saw were randomly sampled from that Deck.

As in the real game Taboo, players took turns. All sessions consisted of 8 Rounds, where each Round involved each Speaker getting one turn. Halfway through the session (after Round 4), participants swapped partners. If they began the session partnered with a friend, they were partnered with a randomly-selected stranger from the opposite team in the second half of the game; if they began the session with a stranger, they were partnered with their friend in the second half. This within-subject manipulation of Partner Type (Friend vs. Stranger) allowed us to control for subject-level differences in NRE production when asking about the effect of Partner Type on NREs.

Partners always sat across from each other (to ensure that the Guesser could not see the Speaker's cards). The experiment itself was presented on a browser screen on a Mac laptop, and was programmed in JsPsych, version 6.0.5 (De Leeuw, 2015). Each card appeared on the screen with the target word bolded, and the five taboo words listed below. During the experiment, one experimenter sat next to the Speaker with the game laptop and indicated via button-press the outcome of each trial (Win vs. Lose vs. Out of time); the other experimenter sat across from the Speaker and monitored them for any taboo words or gestures (which merited a Loss). Both Speakers and Guessers were recorded using a headset microphone during their turn; see the section below for a description of how the audio data was processed and transcribed. The same microphone channel was used for all Speakers in a given session.

Finally, after the session was complete, each participant completed a modified version of the MINT vocabulary test (Gollan et al., 2012), as well as the Unidimensional Relationship Closeness Scale (URCS), a survey designed to assess how "close" each participant felt with their friend (Dibble et al., 2012). They also reported the gender

they identified as, whether or not they had played Taboo before, and their self-rated expertise at Taboo (on a scale from 1 to 5).

Design. The main experimental variable, Partner Type (Friend vs. Stranger), was manipulated within-subject. Sessions were counterbalanced according to whether players began playing with a Friend or a Stranger.

3.2 Audio Transcription and Data Processing

Audio Transcription. The Speaker audio files for each session were first exported to *.wav* format and run through the ReMeeting software (https://remeeting.com/), an Automatic Speech Recognition tool for transcribing audio interview and meeting data. These transcripts were then converted to a *.csv* format, with each line corresponding to a conversational "turn" as segmented by the ReMeeting software.

The resulting transcripts were then analyzed manually for any errors. A team of research assistants listened to the original audio file while reading the transcript; if an error was detected, they added the correct transcription in a new column. The research assistants also identified the time point in each audio file when the game session started (indicated by a tone); this enabled us to align the speech transcripts with the game data recorded by JsPsych. (Note that the dataset includes *both* the original transcription by ReMeeting and the final, manually corrected transcription.)

Data Alignment. First, we identified the time point in each transcript at which the first turn in a session began. Then, for all utterances that occurred between that timestamp and the timestamp of the next turn, we aligned those utterances with the turn in question. Most turns corresponded to multiple utterances, so the individual turn information was repeated on multiple rows of the dataset (for each unique utterance).

Identification and Tagging of Referring Expressions. We then ran each transcribed utterance through spaCy (Honnibal et al., 2020), an open-source Python library for Natural Language Processing. We used spaCy to identify all "base noun phrases" in a given utterance; these are flat phrases that include adjectival modifiers ("the large glass"), but not embedded noun phrases ("[the large glass] on the table").

Then, we assigned a coarse Form Tag to each noun phrase. We expected that if common ground did modulate features of the NREs produced, this should be most observable among Full NPs and 3rd -person pronouns. 3rd -person pronouns (e.g., "he") are typically more ambiguous than a Full NP (e.g., "the dog")—so if common ground licenses more ambiguous NREs (Tily & Piantadosi, 2009), we should observe a higher rate of 3rd -person pronouns among friends than strangers. Further, because Full NPs can vary in their amount of detail (e.g., the number of adjectival modifiers), we hypothesized that common ground might license shorter, less detailed NREs. Beyond tagging the NREs of direct interest to our primary research question (i.e., Full NPs and 3rd -person pronouns), we also sought to provide sufficient information to other researchers interested in potential patterns in how different NREs distribute across a conversational corpus; thus, we also tagged other pronouns (i.e., 2nd -person and 1st -person), and provided Granular Form Tags for a more fine-grained analysis.

Possible Form Tags included:

1. **Full NP**: any phrase beginning with a determiner, possessive pronoun, or quantifier (e.g., "many").
2. **3rd-person pronoun** ("it", "he", "she", etc.).
3. **1st-person pronoun** ("I", "we", etc.).
4. **2nd-person pronoun** ("you", "yourself", etc.).
5. **Other**: a general-purpose category including undetermined or plural NPs ("dogs"), wh-NPs ("which dog"), and proper nouns ("Paris").
6. **Unknown**: a noun phrase identified by spaCy that did not fall into one of these categories.

As noted above, we also assigned a Granular Form Tag to differentiate between some of the categories described above. These differentiated between determined NPs (which we still called Full NPs) and Possessive NPs, along with Plural NPs, wh-NPs, and Proper Nouns. These Granular Form Tags were not used for the planned analyses described below, as we did not have specific hypotheses about how they ought to vary as a function of Partner Type. However, they are still included in the final corpus.

Finally, we included information about the grammatical role a particular noun phrase occupied, in the form of a labeled dependency arc. For example, in the sentence "the dog chased the cat", "the dog" would receive the label *nsubj*, and "the cat" would receive the label *dobj*. We did not have specific hypotheses about how noun phrase properties would change as a function of grammatical role, but we have included this information in the dataset in case other researchers are interested in this question.

In the resulting dataset, each noun phrase is listed on a separate row. Thus, each *turn* corresponds to multiple utterances, and each utterance typically corresponds to one or more noun phrases.

Merging with lexical statistics. For the planned analyses described below, we merged the corpus with datasets containing lexical statistics about the target word: the Brysbaert concreteness norms (Brysbaert et al., 2014), the SUBTLEX English frequency data (Brysbaert et al., 2012), and the Kuperman Age of Acquisition Norms (Kuperman et al., 2012). In the analysis of the behavioral data, we also included another dataset containing a measure of the average distance between each target and the five "Taboo words", where distance was measured as cosine distance between the Glove embeddings (Pennington et al., 2014) for a given target word and a given Taboo word. Glove word embeddings are vectors of real-valued numbers representing a particular word's distributional profile in a given corpus (in this case, the embeddings were trained on Wikipedia and Gigaword 5, a total of 6B tokens). Proximity in vector-space—as measured by the cosine distance between two vectors—is often used as a measure of association or relatedness (Pennington et al., 2014), and is correlated with human judgments of relatedness. (Note that not all target words appeared in these datasets. Thus, in our presentation of the descriptive statistics below, we focus on properties of the corpus *before* merging with these datasets; in the Planned Analysis section, we also list corpus properties after merging with those datasets.)

We focused on concreteness, frequency, and age of acquisition for several reasons. First, although there is considerable debate concerning the underlying mechanisms (Adelman et al., 2006) each variable appears to be correlated with the ease of

word recognition and recall (Brysbaert et al., 2014; Morrison & Ellis, 1995; Allen & Hulme, 2006; Fliessbach et al., 2006; Jessen et al., 2000; Bottini et al., 2021). Thus, these variables were useful both as controls (i.e., to ensure that a potential effect of Partner Type was not statistically confounded with a variable like frequency), and for investigating questions about how they each affected reference resolution in an interactive task. Crucially, our analysis asked about the independent effect of each variable, controlling for the others, allowing us to determine the unique impact of each predictor. Second, past work (Zdrazilova et al., 2018) using a similar methodological paradigm found an effect of concreteness on communicative success. Thus, we sought to replicate this finding, while controlling for the other variables. Finally, we included a measure of the average Glove distance to control for potential differences across cards in how related the target concept was to the Taboo words; we anticipated that variation in this measure could relate to communicative success (e.g., a speaker might find it harder to construct clues for cards with closely related Taboo words).

4 Results

All analyses were performed in R version 3.6.3 (R Core Team, 2020). Mixed effects models were constructed using the *lme4* package (Bates et al., 2015), and nested models were compared using log-likelihood ratio tests. All pre-registered analyses are marked as such and originally described on OSF (https://osf.io/hcya6).

Finally, note that the analyses and descriptive statistics below were calculated from two different datasets. They differed in the number of observations since non-native speakers were excluded from one, and the dataset was merged with information about the target concept's concreteness, frequency, and age of acquisition in one.

Analyses describing solely behavioral results in the absence of any linguistic properties (e.g., communicative success) were run on the original game data, containing 4,377 observations (4,204 after removing non-native speakers; 3,731 after merging with the lexical statistics; and 3,713 after restricting the analysis to target concepts that were Nouns, Adjectives, or Verbs).

Analyses focusing on or including linguistic features (i.e., with a separate row for each identified noun phrase) were run on the final aligned dataset, containing 20,521 observations (19,449 after removing non-native speakers). For any descriptive statistics listed below (e.g., the number of NREs per turn), the dataset used contained 19,449 observations (i.e., all NREs produced by non-native speakers). For any statistical analyses, the dataset was further restricted to target concepts for which we also obtained the requisite lexical statistics (17,219 after merging with the lexical statistics datasets; and 17,107 after restricting the analysis to target words that were Nouns, Adjectives, or Verbs). Both datasets will be made available on OSF (https://osf.io/pxqvb/) and GitHub (https://github.com/seantrott/scarfs).

4.1 Descriptive statistics

Game Data. There were 17 sessions for which we obtained both game data and audio transcriptions. The average number of trials (i.e., cards seen) across sessions

was 247.47 (SD[1] = 51.12, median = 256). There was a relatively even breakdown of which sessions had players begin with a friend (9) vs. a stranger (8). A little over half of participants said they had played Taboo before (34), and slightly less than half (31) said they had not played before; the mean self-rated Taboo expertise was 2.62 (SD = 1.07). The mean Closeness Score was 4.53 (SD = 1.36), with the highest possible score being 7.

In terms of overall communicative success, more than half (55.54%) of trials were Won; 31.9% were Lost (i.e., the Speaker used a taboo word or gesture), and 12.56% were Out of Time. The mean number of cards seen per turn was 8.08 (SD = 1.96, median = 8), and ranged from 4 to 14.

Linguistic Properties. After removing NREs produced by non-native speakers, the total number of noun phrases identified across all 17 sessions was 19,449. The mean number of noun phrases per two-minute turn was 38.29 (SD = 14.84), and ranged from 1 to 106. The mean number of noun phrases per trial (where a trial corresponds to each unique card that a given Speaker saw) was considerably lower (M = 5.73, SD = 4.21), and ranged from 1 to 42.

The most common Form Tag was Full NP (N = 5,408), a category that includes NPs beginning with a determiner (e.g., "the") as well as those beginning with a possessive pronoun (e.g., "my") and quantifiers (e.g., "many"). However, the latter two categories are relatively rare (0.04 and 0.02 of the total) compared to determined NPs (0.22). The full breakdown of noun phrases by Granular Form Tag is depicted in Fig. 1 below.

The mean number of words per noun phrase (including pronouns) was 1.43 (SD = 0.71), and ranged from 1 to 8. Considering *only* the set of Full NPs, the mean number of words was 2.27 (SD = 0.59, median = 2), and ranged from 2 to 8 words (see Fig. 2).

Not surprisingly, some determiners, head nouns, and pronouns were considerably more frequent than others. For Full NPs, the indefinite article "a" constituted 35.1% of all determiners, and "the" constituted 28.5%—the next most frequent determiner was "your", which accounted for only 7.8%. The most frequent head noun was "lot" (4.2%), followed by "word" (3.9%),"thing" (2.9%), "type" (2.5%), and "place" (1.7%). Similarly, the 3rd-person pronoun "it" accounted for 76.3% of all 3rd-person pronouns in the corpus, followed by "they" (17.2%), "them" (3.3%), and "he" (1.9%). These differences are displayed in Fig. 3.

4.2 Statistical analyses

Analysis of behavioral data. We carried out several sets of analyses to ask which variables predicted communicative success on any given trial. There were three possible trial outcomes (Win, Lose, and Out of Time). Our primary interest was in predicting communicative success (Win), so we conducted separate analyses in which success was contrasted either with failing to resolve a referent in time (Out of Time) or using a Taboo word or gesture (Lose). Thus, there were two contrasts of interest: Win vs. Lose, and Win vs. Out of Time. Note that apart from the dependent variable, the

[1] SD = standard deviation.

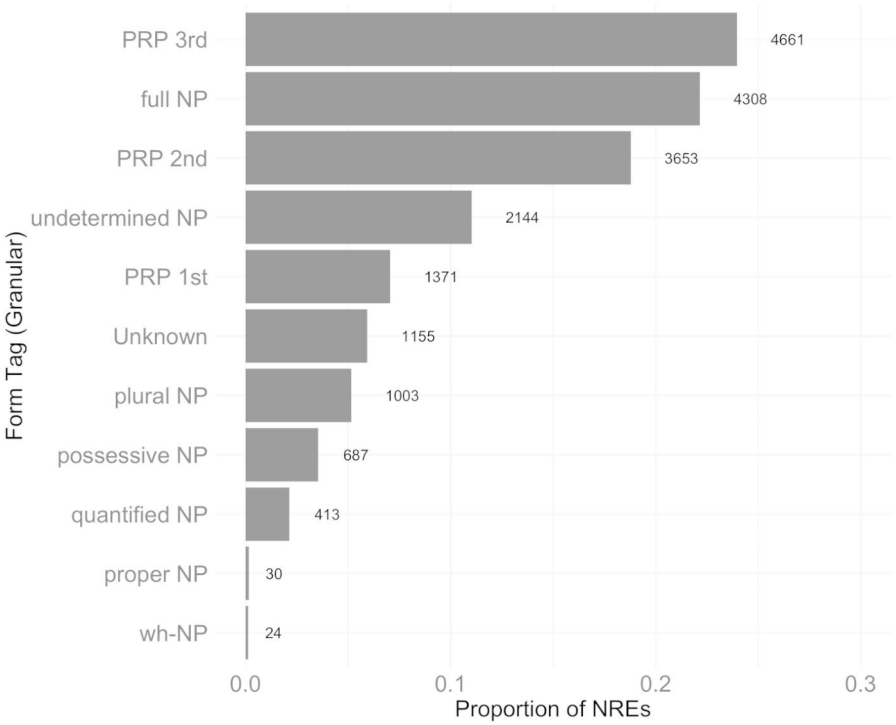

Fig. 1 Proportion (and count) of NREs annotated with each Granular Form Tag (i.e., full NP vs. 1st -person pronoun)

models under consideration were identical (i.e., same fixed and random effects and same model comparisons).

In the first set of analyses, our primary interest was whether either measure of *common ground* played a role in the outcome of a trial (either Win vs. Lose, or Win vs. Out of Time): Partner Type (Friend vs. Stranger) and Trial With Partner (an integer indicating how many trials a Speaker had already played with their partner). We first identified a random effects structure using the approach recommended by Bates et al. (2015). This included: random intercepts for Speaker, Guesser, Session, Deck Color, and Order (i.e., whether the session began with Friends or Strangers), as well as by-Order random slopes for the effect of Partner Type. We also included a fixed effect for MINT Vocabulary Score to adjust for participant-level differences in vocabulary size. Finally, we added an interaction between Trial With Partner and Partner Type, and fixed effects of both, then conducted a series of nested model comparisons.

The interaction did not improve model fit when the target contrast was Win vs. Out of Time ($p > .1$) or Win vs. Lose ($p > .1$). The fixed effect of Partner Type also did not improve model fit for either target contrast ($p > .1$ for both). However, a fixed effect of Trial With Partner did significantly improve model fit for both contrasts: Win vs. Out of Time [$X^2(1) = 8.52,\ p = .004$], and Win vs. Lose [$X^2(1) = 25.5,\ p < .001$]. Interestingly, the effect was in opposite directions in the two comparisons: relative to running Out of Time, a Win outcome became increasingly more likely as the number of turns

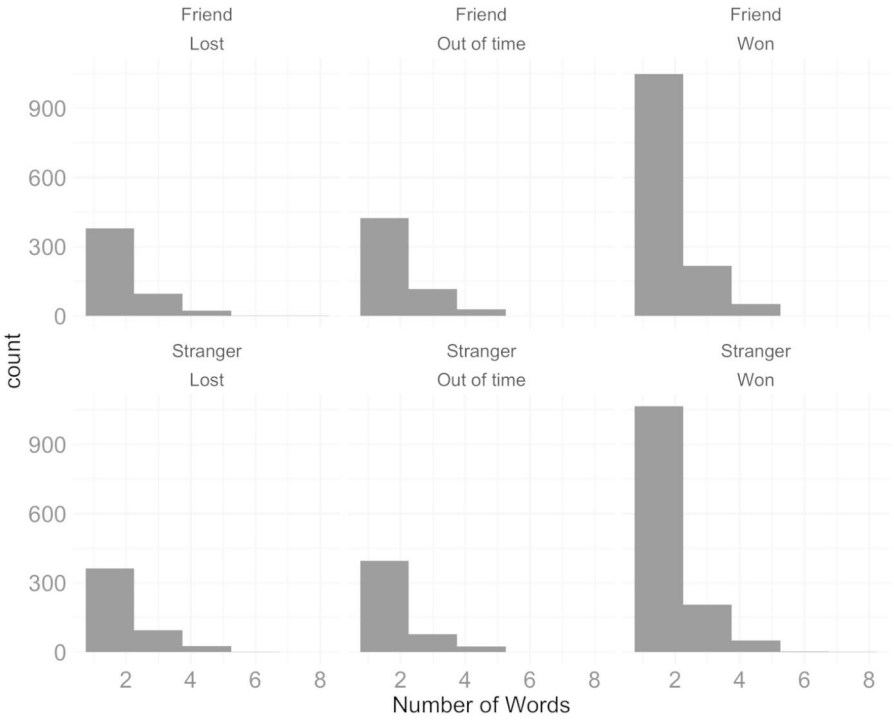

Fig. 2 Number of Words for Full NPs produced for Friends vs. Strangers, broken down by trial outcome (Won vs. Lost vs. Out of Time). As depicted in the graph, there were no differences in the number of words used as a function of Partner Type

each Speaker had with a given partner increased *[ß = 0.02, SE = 0.005]*; but relative to a Loss, the probability of a Win outcome decreased over the course of playing with a given partner *[ß = -0.02, SE = 0.003]*. The former result is depicted in Fig. 4, while the latter result is depicted in Fig. 5.

We also conducted a series of analyses to ask whether linguistic properties of the target word predicted trial outcome. Here, our full model included random intercepts for Speaker, Guesser, Order, Session, and Deck, as well as fixed effects of Trial With Partner, Part of Speech, Concreteness, Log Frequency, and Age of Acquisition. We also included a fixed effect representing the average distance between the target concept and each of its Taboo words, which we called Average Semantic Distance. Because the full model contained covariates for each lexical statistic of interest—Concreteness, Log Frequency, and Age of Acquisition—we could ask about whether each variable explained *independent* variance in communicative success by comparing this full model to a model omitting only that variable.

Model fit was improved by Concreteness when comparing Win vs. Out of Time [$X^2(1) = 21.81, p < .001$] as well as Win vs. Lose [$X^2(1) = 20.67, p < .001$]; in each case, word concreteness was positively correlated with communicative success, replicating a previous effect reported by Zdrazilova et al. (2018). Age of Acquisition was also predictive of Win vs. Out of Time [$X^2(1) = 19.36, p < .001$] and Win vs. Lose

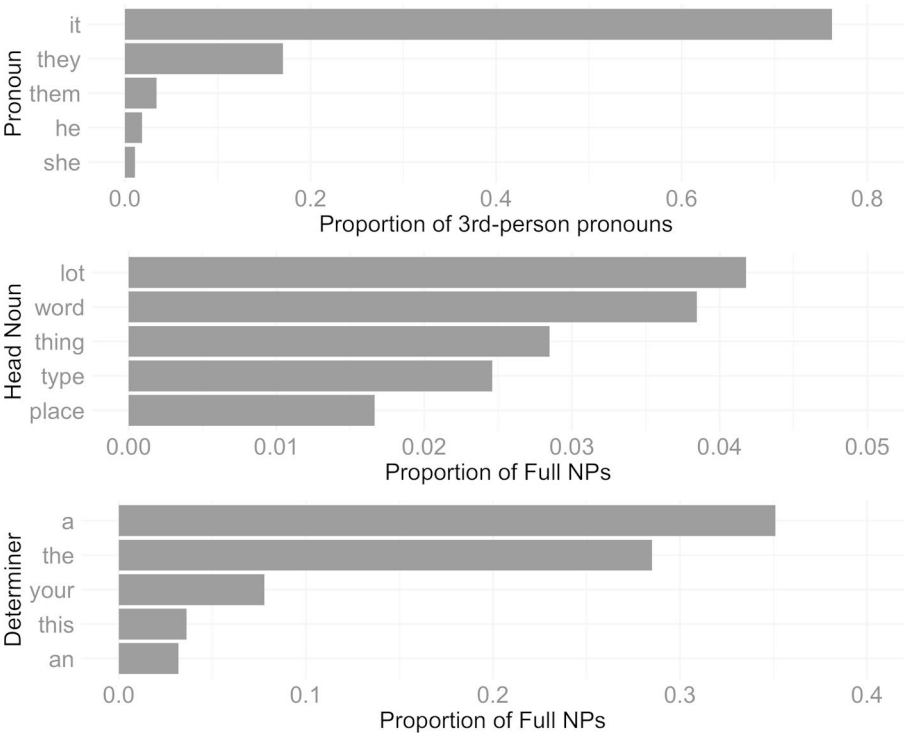

Fig. 3 Proportion of 3rd -person pronouns of each type (Top); Proportion of Full NPs with each head noun (middle); Proportion of Full NPs with each determiner (bottom). Note that this figure only displays the top five entries for each category

[$X^2(1)=13.7$, $p<.001$]; in both cases, words learned later in life were less likely to be communicated successfully (see Fig. 7). Finally, Log Frequency was predictive of Win vs. Out of Time [$X^2(1)=30.37$, $p<.001$] and Win vs. Lose [$X^2(1)=61.85$, $p<.001$]; more frequent words were more likely to be communicated successfully (see Fig. 6). In other words, word frequency, concreteness, and age of acquisition all exhibited independent correlations with the probability of successfully communicating a target concept. There was no independent effect of Average Semantic Distance.

Corpus Analysis. Using the dataset of approximately 20 K noun phrases (17,107 observations after excluding non-native speakers and merging with the lexical statistics), we conducted several pre-registered analyses (https://osf.io/hcya6) to address two questions: (1) Does either measure of common ground influence the Form or Length of referring expressions?; and (2) Does common ground modulate the relationship between referential form and trial outcome?

We addressed the first question in two ways. First, we asked about the *length* of each referring expression. To do this, we considered only NREs that could conceivably vary in length, so we limited our analysis to Full NPs. Thus, considering Full NPs only, we constructed a mixed Poisson model (i.e., with a log link) with Length (Number of Words) as a dependent variable, as well as fixed effects of Trial With Partner, Partner Type, characteristics of the target word (Concreteness, AoA, Log

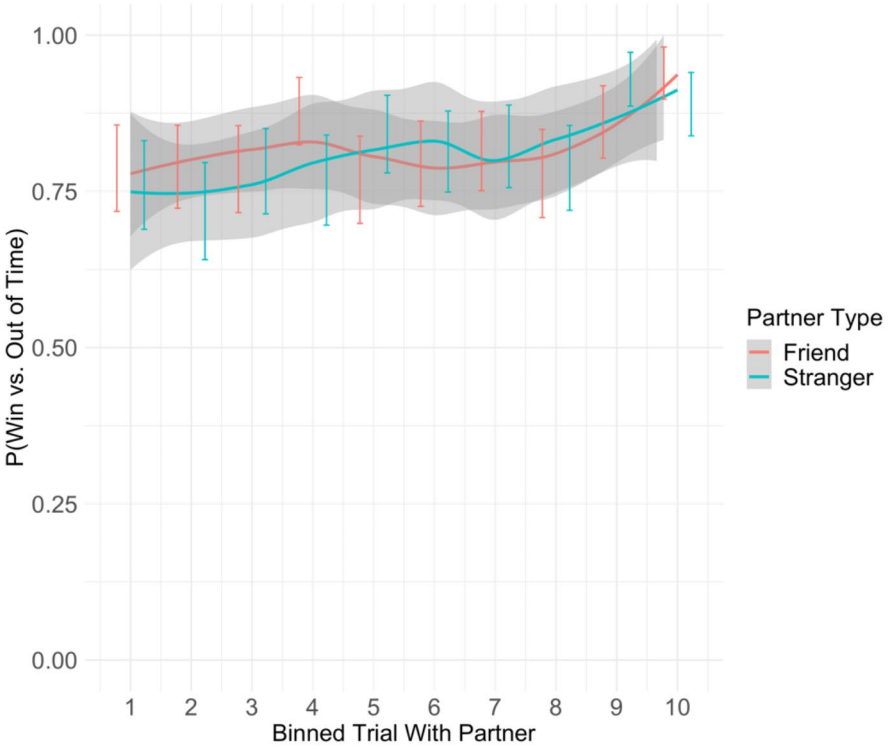

Fig. 4 Proportion of Win outcomes (vs. Out of Time) as a function of each successive trial with a particular partner (binned), broken down by Partner Type. The probability of a Win outcome, relative to running out of time, increased over the course of repeated trials with a partner

Frequency, Part-of-speech), self-reported Taboo Expertise, and MINT score. Random effects were the same as in the behavioral models described above. Model fit was not improved by either Trial With Partner or Partner Type ($p > .1$).

As a second measure, we considered the Referential Form of the NRE produced, e.g., whether friends were more likely to use a 3rd -person pronoun than strangers. Here, we restricted the analysis to two levels of Referential Form variable—Full NP vs. 3rd -person pronoun—because this was the contrast we believed common ground could plausibly play a role in modulating. That is, we did not hypothesize that friends and strangers would differ in the likelihood of using a 1st -person pronoun or 2nd -person pronoun; rather, we hypothesized that if common ground were to play a role in modulating the form used to refer to a referent, it would come into play when referring to referents other than the speaker or addressee. Thus, we constructed a mixed model with a logit link predicting Referential Form (i.e., Full NP vs. 3rd-person pronoun), with identical fixed effects and random effects. Model fit was not improved by Trial With Partner or Partner Type once correcting for multiple comparisons ($p > .1$). Thus, neither measure of common ground was systematically correlated with the length or form of the referring expressions produced.

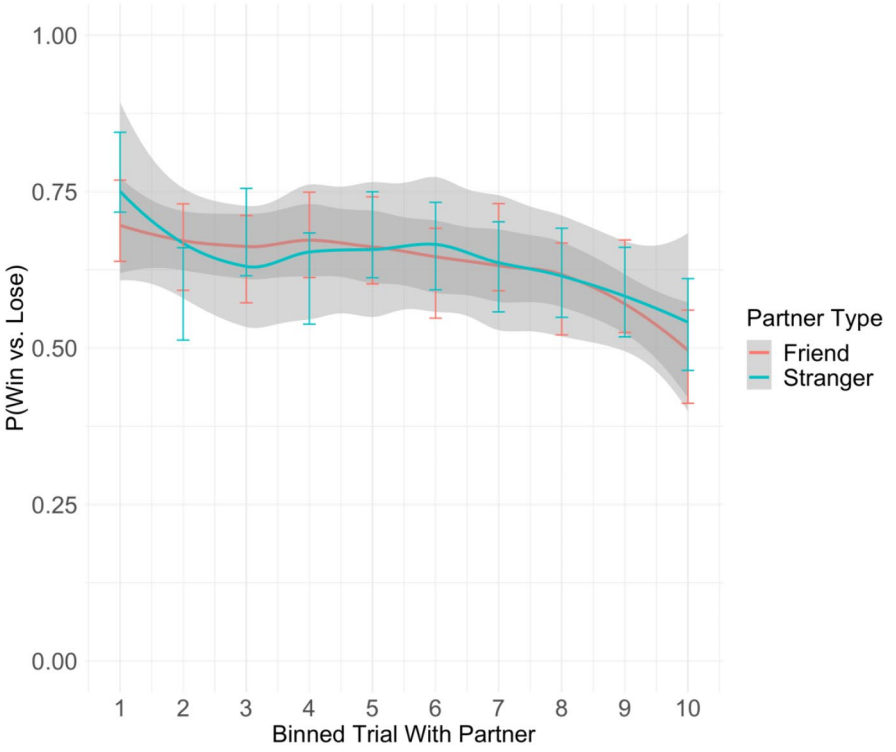

Fig. 5 Proportion of Win outcomes (vs. Lose) as a function of each successive trial with a particular partner (binned), broken down by Partner Type. The probability of a Win outcome, relative to using a taboo word or gesture, decreased over the course of repeated trials with a partner

Our second high-level question was whether the communicative success associated with a particular Referential Form varied as a function of common ground—i.e., whether friends could "get away with" using a 3rd -person pronouns, while strangers could not. In other words, we asked whether there was an interaction between each index of common ground (Pair Type and Trial With Partner) and Referential Form (Full NP vs. 3rd -person pronoun) in predicting communicative success. To address the second question, we constructed two sets of models predicting either Win vs. Out of Time, or Win vs. Lose. In each case, the full model contained four critical interactions: Pair Type and Form, Pair Type and NP Length, Trial With Partner and Form, and Trial With Partner and NP Length. Each model also contained lexical statistics controlling for features of the target word (Concreteness, Age of Acquisition, Log Frequency, Part of Speech), the MINT score of the Speaker, and random intercepts for Speaker, Guesser, Session, Deck, and Order. For each dependent variable, we then compared this full model to a series of four reduced models, omitting each interaction in turn. After correcting for multiple comparisons, none of the variables emerged as significant (all $p > .1$).

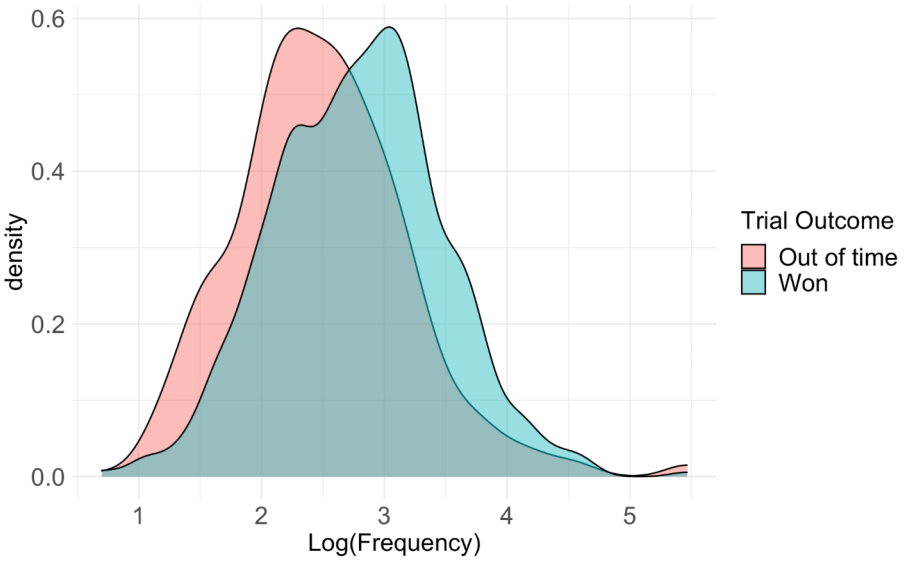

Fig. 6 Log Frequency of the target word by Trial Outcome (Win vs. Out of Time). More frequent target words were more likely to be won than less frequent words

5 Application to NLP and dialogue systems

Both reference resolution and reference generation have longed posed challenges to NLP practitioners (Sukthanker et al., 2020; Zheng et al., 2011; Dale & Reiter, 1995; Williams & Scheutz, 2017). Accordingly, annotated corpora serve an important role in guiding research, particularly for interactive dialogue systems—both as *training data* (e.g., to build systems that display more humanlike dialogue), and as *challenge sets* (e.g., to evaluate how successfully a given system produces referring expressions and resolves intended referents from a referring expression).

The SCARFS Database is annotated with both the intended referent (i.e., the target concept) and the referring expression itself (i.e., the NRE), which makes it well-suited to both applications. Below, we demonstrate its use as a challenge set. The task was relatively straightforward: given a particular *head noun* (e.g., "tooth"), how successfully could a classifier equipped with some representation of the word's meaning resolve the intended *target concept* (e.g., "dentist")? These results are compared against a random baseline and a human benchmark.

We considered two kinds of semantic representation for each word. First, we used Glove word embeddings (Pennington et al., 2014). Second, we used the Lancaster Sensorimotor Norms (Lynott et al., 2020). Like Glove, the Lancaster Sensorimotor Norms contain vector representations for each word. Unlike Glove, the vectors reflect human judgments about the salient sensorimotor properties of those words—that is, to what extent a given word is associated with particular sensory modalities (e.g., hearing, vision, etc.) and action effectors (e.g., hand/arm, foot/leg, etc.). This comparison allowed us to ask to what extent successful reference resolution could be achieved by relying on either a word's distributional statistics or its sensorimo-

tor associations, which echoes more general theoretical (Andrews et al., 2014) and applied (Kiros et al., 2018; Bender & Koller, 2020) debates about where linguistic knowledge comes from. That is, which form of knowledge representation is more successful for this reference resolution task? (In theory, alternative semantic representations could be tested as well.)

5.1 Methods

Data. For this demonstration, we focused on the relationship between head nouns and target concepts. Thus, we limited our analysis to NREs that took the form of Full NPs. We also limited our analysis to words for which we had both Glove embeddings and Lancaster Sensorimotor Norms. Altogether, this corresponded to a total of 4,168 observations. Each observation consisted of a given target concept (the referent) and the head noun corresponding to the NRE produced as a hint for that concept.

Procedure. Using Glove (Pennington et al., 2014) and the Lancaster Sensorimotor Norms (Lynott et al., 2020), we obtained word embeddings for each target concept and each hint's head noun. We then asked whether the word embeddings for the head noun (e.g., "ocean") could be used to infer the corresponding target concept for which that hint had been generated (e.g., "sushi").

This inference (or "guessing") process was implemented as a variation on *k*-nearest-neighbors, a common classification technique. Given an embedding (either Glove or Lancaster) for a particular head noun (e.g., "ocean"), we identified the top *k*-nearest neighbors in vector-space, using cosine distance as a measure of proximity. The value of *k* ranged from 1 (corresponding to the top "guess") to 100 (corresponding to the top 100 "guesses"). Importantly, only words corresponding to possible target concepts were considered in the space of potential neighbors; this made guessing much easier, since there were only 415 possible concepts to choose from.[2] A more challenging test would require selecting from the 40,000 words included in the entire Lancaster dataset.

After identifying the top-*k* nearest neighbors, we asked whether any of those neighbors corresponded to the correct word. For the hint "ocean", the top 5 Glove neighbors included: "island", "earth", "tropical", "beach", and "lake". In this case, the top 5 guesses did not include the true target concept, "sushi", so the response would be marked as incorrect.

We implemented this procedure for a range of values of *k*: 1, 5, 20, 40, 60, 80, and 100. We also calculated a baseline corresponding to random chance, i.e., the probability of the top-*k* guesses including the target concept if one were simply guessing randomly. When $k = 100$, the probability of success with random sampling is approximately 24%. Critically, we did this for both embedding spaces—Glove and Lancaster. This allowed us to ask whether the representational structure of one embedding space (e.g., Glove) provided more information about the relationship between target concepts and hints than the other (e.g., Lancaster).

[2] Although there are 471 concepts in the entire dataset, there were only 454 that met the criteria listed above (i.e., the hint used a Full NP, and both the target and the head noun of the hint appeared in Glove).

Fig. 7 Age of Acquisition of the target word by Trial Outcome (Won vs. Out of Time). Words that were learned earlier on average were more likely to be guessed correctly than words that were learned later on average

6 Results

The accuracy rate for both Models, for each value of *k*, is depicted in Fig. 7, along with the random baseline for each value and the human benchmark for these items (56% accuracy). Across both models, Accuracy varied considerably depending on the value of *k*. Smaller values of *k* had lower accuracy (when $k=1$, accuracy was ~3.41%), while larger values had higher accuracy (when $k=100$, accuracy was ~41.44%). This was to be expected, given that a larger value of *k* corresponds to a larger number of guesses.

Likewise, across both models, average performance exceeded the random baseline for all values of *k*. Even at the largest values of *k*, no models reached human performance (56%). It is worth noting that this is despite the task being made considerably easier for the classifier; the classifier only had to sample from the set of possible target concepts, whereas human guessers did not have access to this set. (Of course, human guessers did have access to the complete utterance produced by the speaker (e.g., "What do they sell at Round Table?"), rather than just the NREs in that utterance.)

Finally, among the different Models, it is clear that Glove exhibited better performance than the Lancaster norms, regardless of the value of *k*. Thus, at least on this task, the structure of the Glove representational space was more amenable to discovering the correct associations between hints and target words than the structure of the Lancaster space was. Put another way, distributional statistics derived from a large corpus of word uses were more informative about the relationship between hints and target words than sensorimotor judgments about those words (Fig. 8).

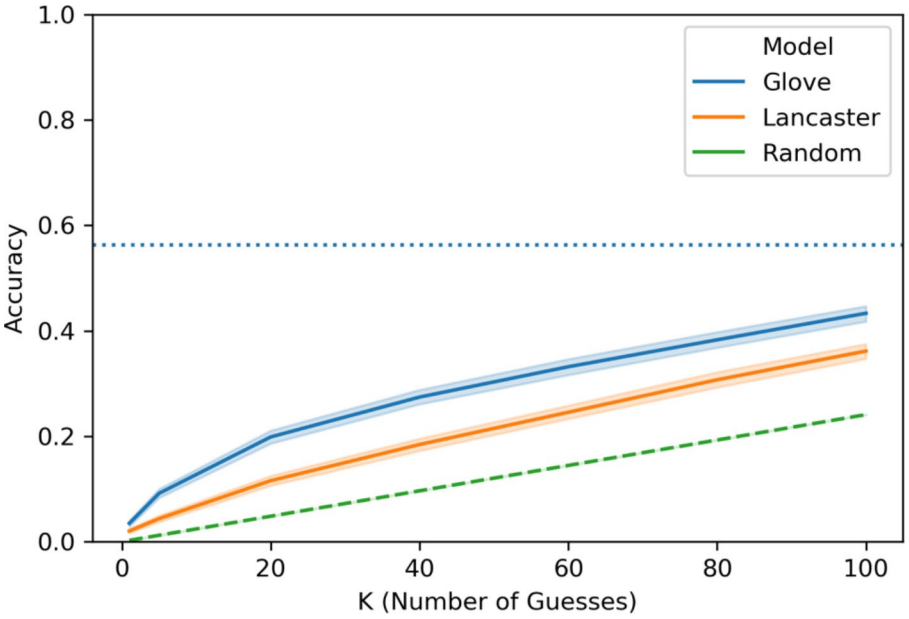

Fig. 8 Accuracy of each Model (Glove and Lancaster) as a function of the number of guesses (K). Accuracy is also compared to the performance of a random baseline (green dashed line) and the human benchmark (blue dotted line). Both Models (Glove and Lancaster) perform better than random chance but are considerably worse than the human benchmark

7 General discussion

Reference plays a central role in communication, but studying it is challenging. On the one hand, researchers can use open-ended dialogue corpora to study how properties of referring expressions vary across communicative contexts. However, these corpora are typically not annotated for whether any given referring expression was "successful" (i.e., whether the comprehender resolved the reference); additionally, the referents themselves are usually not controlled, making it hard to answer questions about causality. The other approach—psycholinguistic experiments conducted in a laboratory setting—affords greater experimental control, but this comes at the cost of restricting the number of referents under consideration.

Here, we introduced the SCARFS Database. The corpus is both relatively large (over 19,000 NREs, and 471 possible target concepts), and also well-controlled and richly annotated. Each act of reference is linked to the target concept a given speaker is attempting to communicate—i.e., the "ground truth". This is useful both for experimental control, and for asking questions about how properties of the concept itself affect the lexical items and grammatical constructions a speaker employs to communicate that referent; this is discussed in greater detail below.

In addition to the "ground truth" concept, each referring expression is annotated with the following information: (1) whether a given act of was successful, or whether the comprehender ran out of time before resolving the referent; (2) the nature of the relationship between the speaker and comprehender (i.e., friend vs. stranger); (3)

how long the speaker has been playing with this particular comprehender; (4) the original dialogue turn in which the referring expression was found; and (5) the grammatical role (e.g., *nsubj* or *dobj*) of the noun phrase in question.

We also carried out several statistical analyses on the corpus. We found no difference in the length or form of referring expressions or in the probability of communicative success across friends and strangers. This finding contradicts the hypothesis that pre-existing common ground ought to facilitate communicative success, as well as the hypothesis that speakers refer differently as a function of their relationship to the addressee; at the same time, it is consistent with past work (Pollmann & Krahmer, 2018; Schober & Carstensen, 2010) suggesting that this type of common ground does not always facilitate efficient, accurate communication. See the *Supplemental Materials* for more discussion of this issue. However, our analysis of behavioral data did replicate a previous finding (Zdrazilova et al., 2018): communication about concrete concepts was more successful than communication about abstract concepts. Other analyses also revealed independent effects of word frequency (communication was more successful for more frequent words) and age of acquisition (communication was more successful for earlier-learned words). Surprisingly, we also found that the probability of using a "Taboo" word or gesture increased as two participants played together for a longer time; future work could investigate this finding further to see whether it is primarily an effect of game "fatigue" or growing comfort with an interlocutor.

Finally, we demonstrated that the SCARFS Database could be of use for NLP researchers working on reference resolution and dialogue systems. Because each act of reference is annotated for its communicative success, researchers can compare the performance of a computational model to this human benchmark. In our case, we found that models using two (relatively simple) representations of a word's meaning out-performed a random baseline, but were considerably lower than human performance—leaving ample room for improvement. While this demonstration was relatively limited in scope (e.g., we restricted the analysis to hints using a Full NP), it serves as a proof-of-concept of the corpus's potential utility to applied questions.

8 Future work

Beyond the analyses carried out here, we hope the SCARFS Database can be used by the broader research community, from linguists to NLP researchers, to address a number of diverse questions.

First, as noted above, each referring expression is annotated with the target concept a speaker is trying to convey. The number of target concepts is relatively large (471) and includes both concrete words like "cactus" and relatively more abstract words like "coincidence" and "honesty."[3] Thus, researchers can ask questions about how particular properties of the referent correlate with a speaker's choice of referring expression, and whether this in turn impacts communicative success. For example,

[3] Although the corpus skews concrete (M = 4.22, median = 4.59), there are still 20 words at least as abstract as "freedom", which is often used as the canonical example of an abstract word (Lupyan & Winter, 2018).

we found that communication about concrete concepts was more successful; is this because speakers use systematically different constructions or lexical items when communicating about concrete concepts, or are comprehenders simply better at identifying concrete referents? Of course, concreteness is only one way of carving up the space—researchers could also ask questions about the kind of referent (e.g., is it an entity, event, or property), or make even more fine-grained divisions (e.g., animate vs. inanimate, human vs. non-human).

Second, the corpus could be used to identify phenomena of interest to conversation analysts, such as self-initiated repairs (Schegloff et al., 1977), disfluencies, and the management of epistemic territory (Heritage, 2012; Bristol & Rossano, 2020). In principle, researchers could ask whether disfluencies (or self-repairs, etc.) distribute differently across friends and strangers, whether speakers are more likely to produce disfluencies when communicating about infrequent or abstract concepts, and whether the use of a disfluency correlates with communicative success.

Similarly, because each pair played for multiple turns, and each turn contained a number of distinct cards, one could ask whether members within each pair exhibit *linguistic alignment*—i.e., converging on the use of similar lexical items or syntactic constructions (Pickering & Garrod, 2004)—and whether the probability of alignment varies across friends and strangers. Other researchers (Pickering & Garrod, 2004) have suggested that alignment underlies successful communication. Thus, given some measure of the degree of alignment between two partners, one could ask whether that measure correlates with the probability of success.

Finally, the Taboo game places unique constraints on reference: speakers must communicate a target concept without using that word, or five of the most related words. This renders the game unnatural in many respects—speakers wishing to communicate the concept "coincidence" would typically just say that word—but it also suggests an interesting application of the words that speakers *do* produce. A common method for measuring the relatedness between two words or concepts is using free association (Nelson et al., 1998): participants are given a word, and asked to produce the first words that come to mind. The proportion of times a given word is produced can be used as a proxy for its relatedness to the target concept. One potential limitation to this approach is that participants might consistently "sample" the most frequent, related words, making it hard to measure the relatedness between a target concept and less frequent or less related words. By forcing speakers to use less typical words to describe a target concept, we can obtain these estimates. For example, the target concept *ache* prohibited participants from using "tooth", "head", "sore", "pain", or "back". Participants in our sessions produced a total of 28 full noun phrases for this concept—of these NPs, 18% used the head noun "body", and 10.7% used the head noun "dentist"; only 3.6% used head nouns like "toe" or "cavity". While there are clearly limitations to measuring relatedness this way, the Taboo methodology could prove a useful complement to existing approaches (Nelon et al., 1998; Hill et al., 2015).

9 Conclusion

We have described a corpus of English NREs, which were produced in the context of a communication game. Each expression is annotated for its formal properties (e.g., length), as well as its communicative outcome (i.e., whether the expression was successfully resolved). We hope the SCARFS Database can help address theoretical questions about how people refer in dyadic communication, and also inform computational models of reference generation and resolution for natural language understanding.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10579-022-09619-y.

Authors' contributions Conceptualization, Methodology, and Writing - Review and Editing [Sean Trott, Benjamin Bergen, Eva Wittenberg]; Data Curation, Methodology, Formal Analysis, Visualization, and Writing - Original Draft [Sean Trott]; Funding Acquisition [Eva Wittenberg]; Supervision [Benjamin Bergen, Eva Wittenberg].

Funding NA.

Data Availability Both the game data and annotated reference data are available on OSF (https://osf.io/pxqvb/) and GitHub (https://github.com/seantrott/scarfs).

Code Availability https://osf.io/pxqvb/ (also GitHub : https://github.com/seantrott/scarfs).

Declarations

Conflicts of interest/Competing interests The authors declare no conflict of interests.

Consent to participate All participants gave written informed consent.

References

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological science, 17*(9), 814–823.

Allen, R., & Hulme, C. (2006). Speech and language processing mechanisms in verbal serial recall. *Journal of Memory and Language*, *55*(1), 64–88

Andrews, M., Frank, S., & Vigliocco, G. (2014). Reconciling embodied and distributional accounts of meaning in language. *Topics in cognitive science*, *6*(3), 359–370

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:https://doi.org/10.18637/jss.v067.i01

Bender, E. M., & Koller, A. (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198)

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning memory and cognition*, *22*(6), 1482

Bristol, R., & Rossano, F. (2020). Epistemic trespassing and disagreement. *Journal of Memory and Language*, *110*, 104067

Brown-Schmidt, S. (2012). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes*, *27*(1), 62–89

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior research methods*, *44*(4), 991–997

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, *46*(3), 904–911

Canavan, A., Graff, D., & Zipperlen, G. (1997). CallHome American English speech. *Linguistic Data Consortium*

Canavan, A., & Zipperlen, G. (1996). CallFriend American English-non-southern dialect.Linguistic Data Consortium, Philadelphia, *10*(1)

Cirik, V., Berg-Kirkpatrick, T., & Morency, L. P. (2020, July). Refer360∘: A Referring Expression Recognition Dataset in 360: A Referring Expression Recognition Dataset in 360∘ Images Images. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7189–7202)

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication.

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science*, *19*(2), 233–263

De Boer, M., Toni, I., & Willems, R. M. (2013). What drives successful verbal communication? *Frontiers in human neuroscience*, *7*, 622

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods*, *47*(1), 1–12

Dibble, J. L., Levine, T. R., & Park, H. S. (2012). The Unidimensional Relationship Closeness Scale (URCS): reliability and validity evidence for a new measure of relationship closeness. *Psychological assessment*, *24*(3), 565

Di Eugenio, B., Jordan, P. W., & Thomason, R. H., & D MOORE, J. O. H. A. N. N. A. (2000). The agreement process: An empirical investigation of human–human computer-mediated collaborative dialogs.International Journal of Human-Computer Studies, *53*(6),1017–1076

Fliessbach, K., Weis, S., Klaver, P., Elger, C. E., & Weber, B. (2006). The effect of word concreteness on recognition memory. *Neuroimage*, *32*(3), 1413–1421

Foster, M. E., Bard, E. G., Guhe, M., Hill, R. L., Oberlander, J., & Knoll, A. (2008, March). The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 295–302). IEEE

Fussell, S. R., & Krauss, R. M. (1989). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of experimental social psychology*, *25*(3), 203–219

Gatt, A., Van Der Sluis, I., & Van Deemter, K. (2007, June). Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)* (pp. 49–56)

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992, March). SWITCHBOARD: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on* (Vol. 1, pp. 517–520). IEEE Computer Society

Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., & Cera, C. M. (2012). Self-ratings of spoken language dominance: A Multilingual Naming Test (MINT) and preliminary norms for young and aging Spanish–English bilinguals. *Bilingualism: language and cognition*, *15*(3), 594–615

Gundel, J. K., Ntelitheos, D., & Kowalsky, M. (2007). Children's use of referring expressions: some implications for theory of mind. *ZAS Papers in Linguistics*, *48*, 1–21

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*(1), 43–61

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020). Characterizing the dynamics of learning in repeated reference games.Cognitive Science, 44(6), e12845

Heritage, J. (2012). Epistemics in action: Action formation and territories of knowledge. *Research on Language & Social Interaction*, *45*(1), 1–29

Honnibal, M., Ines, M., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. *Zenodo* (https://doi.org/10.5281/zenodo.1212303)

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*(1), 91–117

Hu, Z., Tree, J. E. F., & Walker, M. (2018, July). Modeling linguistic and personality adaptation for natural language generation. In *Proceedings of the 19th annual SIGdial meeting on discourse and dialogue* (pp. 20–31)

Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of experimental psychology: general*, *116*(1), 26

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., & Wooters, C. (2003, April). The ICSI meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*. (Vol. 1, pp. I-I). IEEE

Jessen, F., Heun, R., Erb, M., Granath, D. O., Klose, U., Papassotiropoulos, A., & Grodd, W. (2000). The concreteness effect: Evidence for dual coding and context availability. *Brain and language*, *74*(1), 103–112

Keysar, B., Barr, D. J., & Horton, W. S. (1998). The egocentric basis of language use: Insights from a processing approach. *Current directions in psychological science*, *7*(2), 46–49

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38

Kiros, J., Chan, W., & Hinton, G. (2018, July). Illustrative language understanding: Large-scale visual grounding with image search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 922–933)

Kunze, L., Williams, T., Hawes, N., & Scheutz, M. (2017, October). Spatial Referring Expression Generation for HRI: Algorithms and Evaluation Framework. In *AAAI Fall Symposia* (pp. 27–35)

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior research methods*, *44*(4), 978–990

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, *52*(3), 1271–1291

Lupyan, G., & Winter, B. (2018). Language is more abstract than you think, or, why aren't languages more iconic? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752), https://doi.org/10.1098/rstb.2017.0137

Ma, Y., Nguyen, K. L., Xing, F. Z., & Cambria, E. (2020). A survey on empathetic dialogue systems. *Information Fusion*, *64*, 50–70

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. transcription format and programs* (1 vol.). Psychology Press

Morrison, C. M., & Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of experimental psychology: learning Memory and cognition*, *21*(1), 116

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/

Niekrasz, J., & Moore, J. D. (2010, July). Annotating participant reference in English spoken conversation. In *Proceedings of the Fourth Linguistic Annotation Workshop* (pp. 256–264)

Orita, N., Vornov, E., Feldman, N., & Daumé, I. I. I. (2015, July). H. Why discourse affects speakers' choice of referring expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1639–1649)

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543)

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, *27*(2), 169–190

Pollmann, M. M., & Krahmer, E. J. (2018). How do friends and strangers play the game taboo? A study of accuracy, efficiency, motivation, and the use of shared knowledge. *Journal of language and social psychology*, *37*(4), 497–517

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*(2), 361–382

Schegloff, E. A. (1996). Some practices for referring to persons in talk-in-interaction: A partial sketch of a systematics. *Typological studies in language*, *33*, 437–486

Schober, M. F., & Carstensen, L. L. (2010). Does being together for years help comprehension? *Expressing oneself/expressing one's self: Communication, cognition, language, and identity*, 107–124

Shintel, H., & Keysar, B. (2009). Less is more: A minimalist account of joint action in communication. *Topics in Cognitive Science*, *1*(2), 260–273

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C., & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, *26*(3), 339–373

Sukthanker, R., Poria, S., Cambria, E., & Thirunavukarasu, R. (2020). Anaphora and coreference resolution: A review. *Information Fusion*, *59*, 139–162

Takenobu, T., Ryu, I., Asuka, T., & Naoko, K. (2012). The REX corpora: A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues. In *Proceedings of LREC* (pp. 422–429)

Tily, H., & Piantadosi, S. (2009, July). Refer efficiently: Use less informative expressions for more predictable meanings. In *Proceedings of the workshop on the production of referring expressions: Bridging the gap between computational and empirical approaches to reference*

Viethen, J., & Dale, R. (2006, July). Algorithms for generating referring expressions: do they do what people do?. In *Proceedings of the fourth international natural language generation conference* (pp. 63–70)

Weischedel, R., Pradhan, S., Ramshaw, L., & Micciulla, L. (2008). *Ontonotes release 2.0*. Linguistic Data Consortium

Williams, T., & Scheutz, M. (2017, September). Referring expression generation under uncertainty: Algorithm and evaluation framework. In *Proceedings of the 10th International Conference on Natural Language Generation* (pp. 75–84)

Willems, R. M., De Boer, M., De Ruiter, J. P., Noordzij, M. L., Hagoort, P., & Toni, I. (2010). A dissociation between linguistic and communicative abilities in the human brain. *Psychological Science*, *21*(1), 8–14

Winters, J., Kirby, S., & Smith, K. (2018). Contextual predictability shapes signal autonomy. *Cognition*, *176*, 15–30

Wheatley, B., Kaneko, M., & Kobayashi, M. (1996). *CALLHOME Japanese Transcripts LDC96T18. Web Download*. Philadelphia: Linguistic Data Consortium

Zdrazilova, L., Sidhu, D. M., & Pexman, P. M. (2018). Communicating abstract meaning: concepts revealed in words and gestures. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752), 20170138

Zheng, J., Chapman, W. W., Crowley, R. S., & Savova, G. K. (2011). Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of biomedical informatics*, *44*(6), 1113–1122