There is a power law of joint communicative effort and it reflects communicative work

Sara Bögels[1,2], Tianyi Li[3], Marlou Rasenberg[4], Lotte Eijk[5], Ivan Toni[1], Wim Pouw[1]

1. Donders Institute for Brain, Cognition and Behavior, Nijmegen
2. Centre for Language Studies, Nijmegen
3. Department of Psychology and Neuroscience, Boston College, Chestnut Hill, Massachusetts, USA
4. Meertens Instituut, KNAW, Amsterdam, The Netherlands
5. Department of Psychology, University of York, United Kingdom

Author Note

Funding: This work was supported by the Dutch Research Council (NWO) Gravitation Grant (024.001.006) to the Language in Interaction Consortium. Wim Pouw is funded by a VENI grant (VI.Veni.201G.047) awarded by the Dutch Research Council (NWO).

Acknowledgements: We thank the CABB team for their discussions on aspects of this work, especially Judith Holler, Asli Ozyurek, Esam Ghaleb, Raquel Fernandez, and Mark Dingemanse.

Contact: sara.boegels@ru.nl, wim.pouw@donders.ru.nl

Open data: Unidentifiable data and analysis code can be found at https://github.com/WimPouw/putting_in_the_effort_CABB.

Abstract

A drive towards efficiency seems to regulate communicative processes and ultimately language change. In line with efficiency principles, signed, spoken, and/or gestural utterances tend to reduce in overall effort over repeated referrals in referential tasks. Such reduction is often studied in individuals, using a single communicative modality. Here we seek to understand reduction of communicative effort in its natural communicative environment, i.e. during multimodal and collaborative face-to-face dialogues about displaced referents. We ascertain that the reduction in joint effort over repeated referrals actually follows a negative power relationship. This reduction in communicative effort is multimodal, occurring across gesture, speech, prosody, and turn taking, and it is interactive, based on joint effort. The effect is robust, being confirmed through a reanalysis of published datasets about (individual) effort reduction. Crucially, the effect is also communicatively relevant. The coefficient of the power relationship predicts change and convergence in interlocutors' conceptualizations of the communicative referents. Assuming that the coefficient of the negative power relationship reflects how well effort translates into mutual understanding - a process we call communicative *work* - we suggest that the power function captures a complementary strategy of increasing initial exploration and applying efficient selection for an effective joint conceptualization of referents. The current report invites linguistic theory, agent-based modeling, and experimental psychological inquiries into understanding the general principles of what could amount to a 'power law of joint communicative work'.

Keywords: reduction, conventionalization, negative power law, efficiency, effort, communication

Introduction

Imagine playing a referential communication game where you describe figures or concepts to another person using all the different modes of communication available to you and drawing on feedback from the other person. Over multiple rounds you, as the director, need to 'direct' the matcher to find the right referent through words and gestures. The matcher contributes too, for example through multimodal backchannels and requests for clarification. A well-known finding in these types of 'director-matcher' tasks is that the effort directors invest in their gestural, signed, and/or spoken utterances to refer to some object or concept tends to reduce over the number of referrals (e.g., over the number of task rounds). While a director might describe or depict a concept in an elaborate way when presented for the first time, on later occasions the utterance used to refer to the same meaning becomes much more stylized, while the matcher maintains the ability to recognize the meaning, in part because of a built common ground.

This general reduction of effort has been found in a range of situations (see Figure 3 in *Results* for an illustration), but it has primarily been studied unimodally and by focusing on the effort put in by single individuals (i.e., directors). For example, repeated references contain fewer words or less speech, in written and spoken discourse (Arnold et al., 2013; Givón, 1983; Marslen-Wilson et al., 1982), interactive story-telling (Arnold et al., 2013; Givón, 1983; Marslen-Wilson et al., 1982), and reference games (Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986; de Ruiter et al., 2012; Hawkins et al., 2020; Hoetjes et al., 2015; Holler et al., 2011). Repeated words or phrases are shorter and/or acoustically reduced (Clark & Haviland, 1977; Fowler, 1988; Fowler & Housum, 1987; Holler et al., 2022). Moreover, gestures in interaction are less likely to occur in repeated references (Alibali et al., 2014; Jacobs & Garnham, 2007; Levy & McNeill, 1992). When they do occur, they are generally shorter, smaller, and/or less precise (Bangerter et al., 2020; Galati & Brennan, 2014; Gerwing & Bavelas, 2004; Hoetjes et al., 2015; Holler et al., 2011, 2022; Vajrabhaya & Pederson, 2018). A similar effect is found in iterated learning tasks, where references are evolving over multiple 'generations', or different sets of interactants, for both verbal (Hawkins, Franke, et al., 2023) and gestural references (Motamedi et al., 2019; Pouw et al., 2021). In contrast, if a matcher formulates repeated references for a different person each time, no reduction of effort is found (Hupet & Chantraine, 1992). What is the nature of this reduction? Can this reduction process be related to general patterns in human cognitive and linguistic processes?

The reduction of effort may be rooted in communicative efficiency, or the principle of least effort (Zipf, 1949), optimizing how effort translates into communicative effects (Gibson et al., 2019; Levshina & Moran, 2021). This relates to Grice's maxim of quantity: say as much as needed, but not more (Grice, 1975). Apart from efficiency of individual communicative utterances, interactants have been shown to minimize *joint* effort as well (known as the principle of least collaborative effort; Clark & Schaefer, 1987; Clark &

Wilkes-Gibbs, 1986), for example, by an optimal division of labor in repair sequences (Rasenberg et al., 2022). In repeated referential games there seems to be an interactive *selection process* manifesting as a reduction or compression of utterances to reach similar referential effects. For example, in research using a Tangram task, where geometrical shapes need to be communicated using descriptions, it was shown that over repeated references of six rounds, more elaborate descriptions were reduced by selecting out so-called 'closed class' parts of speech like pronouns, conjunctions, and determiners. However, 'open class' parts of speech, like nouns and verbs survived and were reused more often for repeated referrals, arguably due to their stronger distinctiveness (Hawkins et al., 2020). Or in the case of a director-matcher game based on line drawings, it was shown that over repeated referrals, drawings reduced in a way that distinctive elements in the drawing survived over repetitions (Hawkins, Sano, et al., 2023). As such, the reduction of communicative effort appears to stem from (parts of) earlier expressions interactively selected for re-use.

Here, we study the dynamics of the reduction in communicative effort, and how it relates to communicative outcomes, in the context of multimodal and collaborative face-to-face dialogues. In two studies, we measured joint multimodal effort by quantifying the combined amount of speech (using acoustic tracking), amount of manual movement as a proxy for gesture (using motion tracking), degree of prosodic modulation (using acoustic tracking), and amount of turn taking (using annotations). In a first study we obtained a particular curvilinear relationship that described the reduction of interactive, multimodal effort.

Curvilinear relationships, including power or logarithmic functions, are ubiquitous phenomena in nature and have been known as objects of psychological study of perception, learning, and problem solving as early as Weber (Weber, 1978), Ebbinghaus (Ebbinghaus, 1885), and Thorndike (Thorndike, 1898). Weber (Stevens, 1961; Weber, 1978) showed that the minimal degree of difference in two stimuli that are perceptually distinguishable decreases as the magnitude of sensory stimulation of the stimuli decreases (Weber's law[1]). Ebbinghaus (1885) showed that the rate of forgetting of stimuli reduces over time as memory gets consolidated. Thorndike (1898) showed that the time it takes for a cat to escape a puzzle-box decreases over the number of trials. If we look closely at these negative curvilinear relationships, which are all best modeled as a negative power relationship, we see that there is a steep drop off in the dependent variable, but the reduction then 'decelerates', almost converging to a plateau.

[1] Also known as the Weber-Fechner law. Gustav Fechner (1801-1889) formalized Weber's observations, positing that the observations followed a logarithmic function. It is now established that the Weber-Fechner law actually follows a power function (Stevens, 1961), similar to what we observe in this paper for effort reduction.

Why should we care about effort reduction being curvilinear? After all, these patterns may be caused by very different underlying dynamics - e.g., power laws show up in the frequency and magnitude of earthquakes (Bak, 1996) and other frequency (Zipf, 1949) or spectral distributions (1/f noise; Gilden, 2001) as well. First, we care because researchers are often inappropriately statistically modeling these relationships as linear, while this assumption is clearly violated (e.g., in our overview of curvilinearities of previous studies in Figure 3 (in *Results*), all studies used linear models with untransformed data). Second, while it is true that the underlying dynamics of curvilinear patterns do not wear their underlying mechanisms on their sleeves, the particular temporal statistics obtained might still be informative for theorizing about a more general set of constraints that could describe multiple kinds of systems having similar temporal statistics. Consider for example, that in Educational psychology, the 'learning curve' follows a negative power relationship entailing that the rate of learning is generally faster at the beginning but then decelerates as there is less room for improvement (Viering & Loog, 2023). This immediately invites the question of whether the learning curve and Ebbinghaus's forgetting curve might be governed by some mechanism related to retention and consolidation that operates at a higher level of description so that it can explain both characteristically curvilinear curves. Here, we also seek a more general description for the observed curvilinear reduction in joint communicative effort by appealing to the notion of *communicative work*.

We consider communicative work as the process of communicative effort translating into task-relevant communicative outcomes, in analogy with how physical work captures the effect of a force translating into task-relevant displacement (Kauffman, 2019). The task-relevant outcomes in our case are related to the joint conceptualizations of a referent along a series of dialogic turns, leading to selecting of the correct referent by the matcher. In our paradigm we can directly test whether the curvilinearity of joint communicative effort reduction relates to communicative effects (i.e., is reflective of work). In earlier research showing reduction in referential games, the communicative *effect* mostly consisted of the matcher being able to find the referent that the director describes. In many tasks this outcome is around ceiling performance since directors are given unlimited time to describe the referent. In the present study, however, we measured communicative effects in a more gradient manner, namely the *change* in conceptualization of a described object within individuals, as well as the *convergence* of conceptualizations between interactants, as a result of the interaction. We reasoned that the task requires partners to change their original conceptualizations through communicative effort so that their descriptions evoke the referent for both interactants to such an extent that it can be recognized. And crucially, we understand this effort to be controlled jointly and constituted multimodally (de Ruiter et al., 2012; Rasenberg, Özyürek, et al., 2022; Rasenberg, Pouw, et al., 2022), moving beyond most individualistic and unimodal conceptions of communicative effort as reviewed above.

The hypotheses and analyses in the current study have been pre-registered (Bögels et al., 2023). The pre-registration also included a preliminary exploratory report (hereafter referred to as *original exploratory study*), in which we analyzed a dataset (*N* = 19 pairs) of a referential communication game. We found a negative power relationship describing the reduction of communicative effort, that is between communicative effort in several modalities (speech, gestures, interaction, prosody) and number of reference, or rounds in the game. We also found that the slope of the negative power relationship (i.e., the pattern of reduction of effort) could predict communicative effects, in the form of the amount of individual conceptual change from before to after the interaction. In our interpretation, this slope, or the rate of effort reduction within the interaction, thus reflects the communicative work performed. Here we report on the confirmatory (pre-registered) analyses of a very similar, but independent larger dataset and external datasets (hereafter referred to as *current confirmatory study*) - with these analyses we aimed to replicate and further flesh out the original exploratory findings.

To summarize, the confirmatory hypotheses that we focus on in this report are:

1. Effort reductions follow a curvilinear slope, better modeled by a negative power relationship as opposed to a linear relationship.[2]
2. More conceptual change relates to more negative effort slopes over the rounds, interpreted as more or faster conceptual progress within the interaction.

We further investigate if hypothesis 1 can retroactively better explain other research findings too, and whether hypothesis 2 also applies to convergence of conceptualizations between pair members (as convergence is arguably related to task success). These analyses are hereafter referred to as *current exploratory work*.

Note that two other pre-registered hypotheses failed to replicate, such that we no longer forefront them in our report here (but they are fully discussed in our pre-registration and the Supplementary Materials, Section S2).

Methods

For our current confirmatory analyses we analyze 42 pairs of participants from the CABB dataset (Eijk et al., 2022). These pairs consist of 17 all-female pairs, 5 all-male pairs, and 20 mixed-gender pairs. The participants were 22.4 years old on average (SD = 3.02, range = 18-33).

The procedure was such that participants who did not know each other formed pairs in the lab. After giving informed consent, they performed several individual tasks,

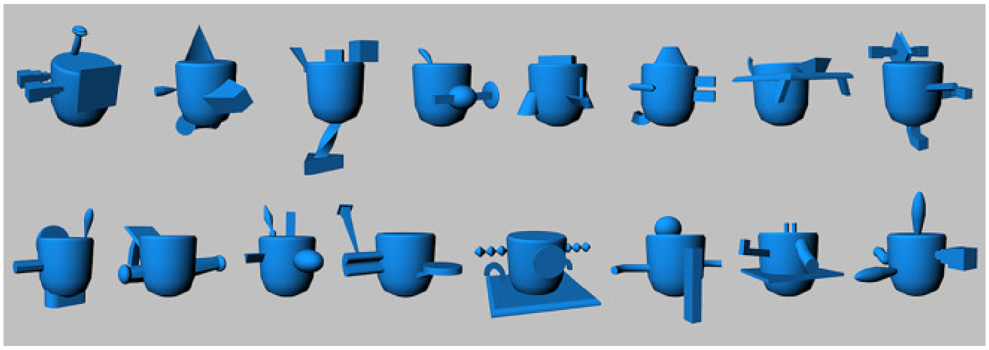
[2] This first hypothesis was not explicitly listed as one of the main hypotheses in our preregistration since it was part of our methodological approach there (Bögels et al., 2023). But we do consider it explicitly as part of our confirmatory analyses in this report. Hypothesis 1 in the preregistration therefore corresponds to hypothesis 2 here.

then interacted together, and then again performed individual tasks. See extended Methods below, and Eijk and colleagues (2022) for a detailed description of the tasks. Key to the current analyses is that participants individually named 16 objects (Fribbles, see Figure 1, based on Barry et al. 2014) using one to three words, both before and after the interaction. They were instructed to make the names informative for their partner. During the interaction, they both saw the 16 Fribbles on their individual computer screens, in different arrangements. In each trial, the *director* described one designated Fribble to the *matcher*, who had to find the Fribble on their own screen, while being allowed to interact with the director. Director and matcher roles switched each trial and all 16 Fribbles were described six times in total (i.e., in six *rounds*). Interactions were audio- and video recorded, and movement measurements were taken from both participants using a Microsoft Kinect system.

Measures of four different types of communicative effort were extracted per trial, and we favored measures that can be automatically obtained from the signals. Speech effort was therefore estimated as the number of peaks within the speech amplitude envelope of the audio file, manual movement effort as the number of movement speed peaks within the Kinect recordings, and prosodic effort as the amount of F0 variation. Interactional effort was based on the number of speaker turn transitions that could only be derived from the transcripts. These four measures were z-normalized for the entire dataset and then averaged per round. We then extracted the intercept and slope of a linear model over the six rounds for each Fribble-pair combination. The intercept of that model reflects the average effort and the slope reflects the decrease in amount of effort over time (i.e., a more negative coefficient indicates high effort at start, and quick drop off). We considered two transformations as input for our linear models, using untransformed and log-transformed communicative effort measures, to see which would lead to a better model fit.

Communicative effect was measured using (1) *conceptual change* (pre-registered; Bögels et al., 2023): estimated as the change in semantic distance between names given to each of the Fribbles, from pre to post interaction per individual, averaged over the two pair members, and (2) *conceptual convergence* (exploratory analyses): estimated as the increase (or decrease) in semantic distance between the Fribbles' names given by each of the two participants of a pair from pre to post interaction. Semantic distances between the Fribbles' names were calculated using word2vec (see Extended Methods for details).

Figure 1. Stimuli (Fribbles) used in the study



Results

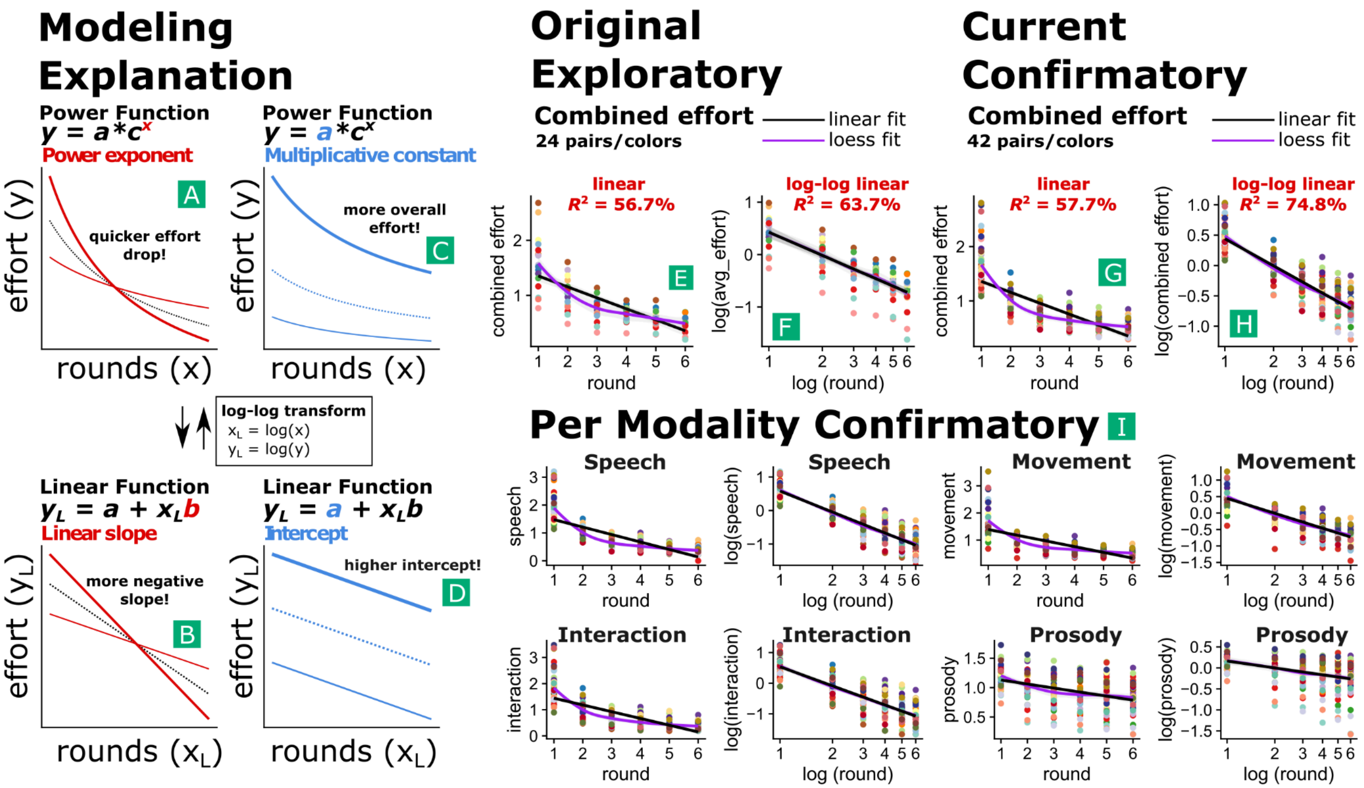
Below, we report the main results. We first report current confirmatory and current exploratory results on the negative power relationship (Hypothesis 1). We then report on current confirmatory and current exploratory results on how that power relationship relates to communicative effects (Hypothesis 2).

Negative power relationship

Current confirmatory analyses

We first assess whether average effort reduction over the six rounds can best be approximated in a linear or curvilinear way. The untransformed model (b = -0.201, t [209] = -19.59, p < .001), and the log-log-transformed model (b = -0.644, t [209] = -41.47, p < .001) both showed a clear negative relation between rounds and effort as can be verified in Figure 2. However, the log-log transformed model (fixed effects + random intercepts) explained 74.8% for fixed effects (89% for entire model) of the variance (based on R^2, using *R*-package MuMIn), while the untransformed model explained 57.7% for fixed effects (62% for fixed effects + random intercepts). Note that when we assess another possible curvilinear relationship called a logarithmic function (where, `y = intercept + log(x)`), we get an explained variance of 70.8% (78% for the entire model) which is considerably lower than the log-log model. We therefore conclude that the overall reduction of communicative effort in this type of task is best approximated by a curvilinear negative power relationship.

Figure 2. Power law fit of combined effort over the rounds versus an untransformed liner fit.



Note. The left panel (Modeling Explanation) provides a description of a power function that can be modeled as a linear relationship when both the independent and the dependent variable are log-transformed (i.e., log-log). In a linear model based on log-log data the slope coefficient reflects the power exponent of the power function, where lower and higher values are shown in A (untransformed) and B (transformed). The intercepts of a linear model on transformed (D) data reflect the multiplicative constant of a power function of the untransformed data (C). On the right panels we show the original exploratory results on the smaller dataset (*N* dyads = 19, middle panel) and the current confirmatory results on the larger dataset (*N* dyads = 42, right panel) which reflect that effort combined (separated out per modality in I) reduces over time. Speech indicates the number of peaks in the amplitude envelope, movement the number of peaks in hand kinematics, interaction the number of turn transitions, and prosody the range in F0. As can be seen in both the current confirmatory and original exploratory analyses we get a clear linear patterning under a log-log transform (F & H) while this is not the case for untransformed data (E & G), that is, the moving average [loess, purple line] does not fit a linear slope well, while it does in the log-log plots.

Exploratory analyses of other datasets

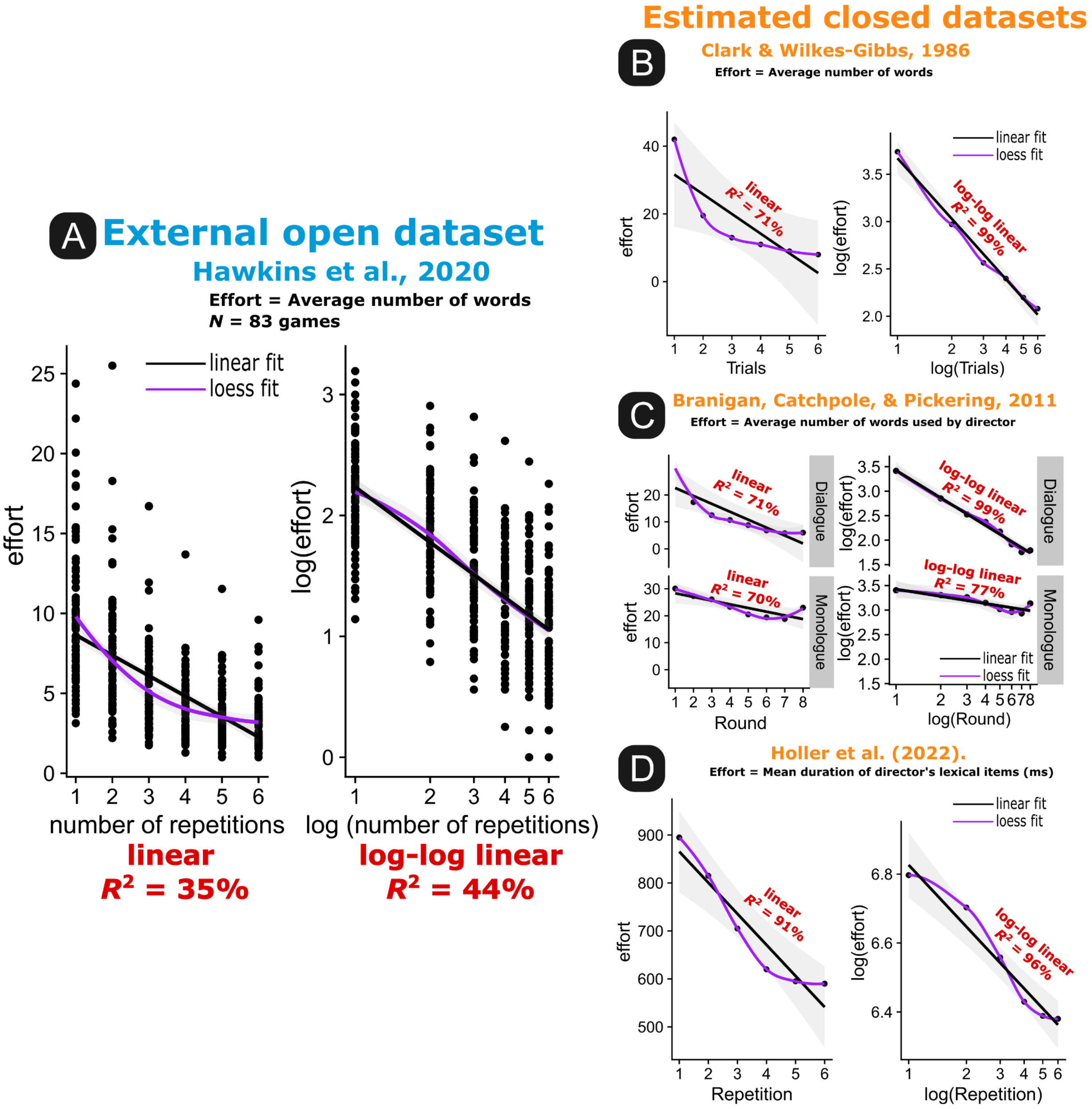
Though we obtain evidence for a power law relationship in a confirmatory fashion within the context of our paradigm, one may wonder whether the power law is a common temporal characterization in repeated referential games in general. Fortunately, the publicly shared data by Hawkins et al. (2020) on a repeated reference game via chat on 12 tangram stimuli (*N* = 83 pairs), allow us the presence of a negative power slope.

We reanalyzed their data (Figure 3, panel A) using a similar mixed regression model predicting overall mean number of words by the director per number of repeating referrals for each pair playing a game (*N* = 83 pairs), with pairs as random intercept. The untransformed linear mixed regression model [b = -1.276, t [414] = -21.84, p < .001] showed 35% explained variance for fixed effects (63% for the entire model). The log-log transformed linear mixed regression model [b = -0.659, t [414] = -30.91, p < .001], showed 44% explained variance for fixed effects, (77% for the entire model). Interestingly, the power law coefficient of -0.6 is similar to our confirmatory results with the CABB dataset presented above. Note, that for another curvilinear relationship called a logarithmic function (where, `y = intercept + log(x)`) we find a lower amount of explained variance 39.4% (68.5%) as compared to the negative power function.

We also performed a weaker set of analyses on other external paradigms that do further support the generalizability of our results that effort reduction follows a curvilinear relation. When reviewing the literature we found several director-matcher type communication studies that also appeared to show curvilinear type of effort reduction (based on (visual) inspection of figures or reported descriptives). The data for these studies are not open, but we can of course extract the summary statistics either from the figures themselves (Clark & Wilkes-Gibbs, 1986; Holler et al., 2022) or from the reported descriptives (Branigan et al., 2011). We can then run simple linear models on untransformed and log-log transformed values to see if the data seem to pattern in a certain way. In all three datasets analyzed in this way, we find clearly that a log-log transform leads to more explained variance[3] (see Figure 3, panels B-D).

[3] Given the few data points for the closed datasets we will not compare different curvilinear relationships as we did for the current confirmatory and the external open dataset.

Figure 3. External datasets showing similar curvilinear relationships.



A External open dataset

Hawkins et al., 2020

Effort = Average number of words
N = 83 games

Estimated closed datasets

B Clark & Wilkes-Gibbs, 1986

Effort = Average number of words

C Branigan, Catchpole, & Pickering, 2011

Effort = Average number of words used by director

D Holler et al. (2022).

Effort = Mean duration of director's lexical items (ms)

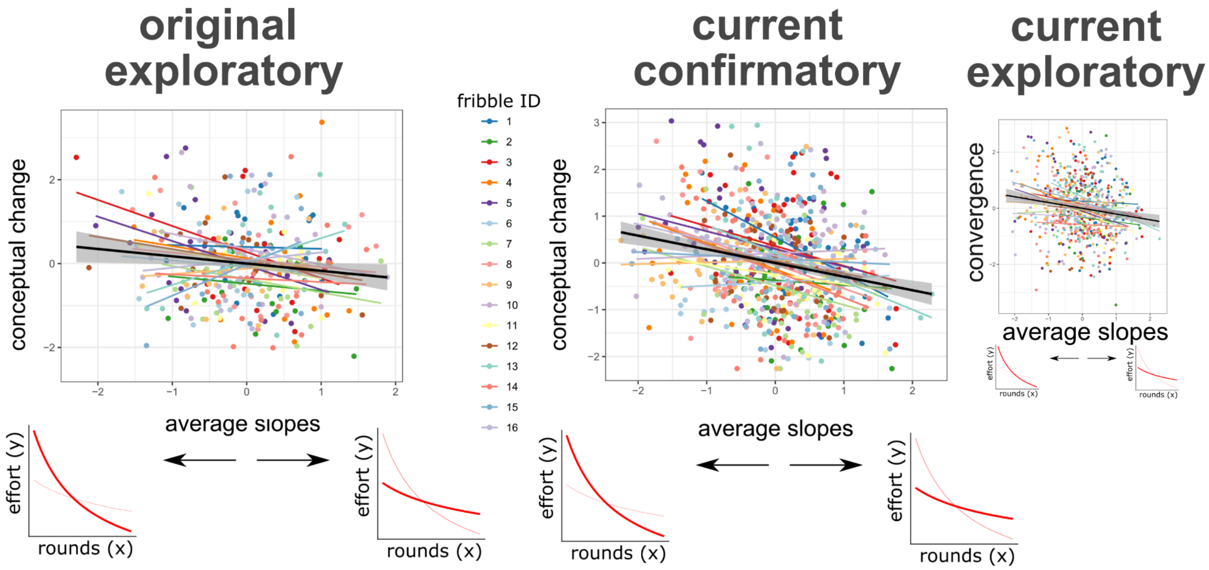
Note. The graphs in all four panels show the untransformed linear fit (left) and the power-law fit (right) to the data of four different previous studies in which pairs of participants played a referential communication game on 12 tangram figures with a fixed director. Panel A shows an open dataset of chat communication in 6 rounds from Hawkins et al. (2020) on 83 pairs of participants. The right panels use only aggregated data (in the absence of availability of the full dataset) based on descriptive data or figures. Panel B: Clark & Wilkes-Gibbs (1983, Figure 2): speech communication without visual contact by 6 pairs, 6 rounds. Panel C: Branigan et al. (2011, Table 2), speech communication without visual contact by 12 pairs (6 for monologue in which the matcher does not speak, 6 for dialogue), 8 rounds. Panel D: Holler et al. (2022, Figure 11), face-to-face communication by 8 pairs, 6 rounds.

Power coefficient and conceptual change

Confirmatory results

Next, we assess whether the obtained negative power relationship relates to conceptual change. A linear mixed-effects model with conceptual change as outcome measure, pair and Fribble as random intercepts[4], and average slope over all four effort measures as fixed effect, showed an effect of slope (β = -0.286; SE = 0.053; t(616.9) = -5.41; p < .0001), which indicates that the steeper (more negative) the slope and thus the more and/or faster the effort decreases over rounds, the more change takes place in Fribbles' conceptualizations (see Figure 4). In our interpretation, this means that more decrease of effort over the rounds (more negative slopes), indicates that more (or faster) conceptual progress has been made, thereby relating to the conceptual change that has resulted. Hypothesis 2 can thus be confirmed.

Figure 4. The effect of average slope coefficient on individual conceptual change and pairwise conceptual convergence from pre to post



Note. The two leftmost graphs depict the relation between average slopes (more negative towards more positive, x-axis) and the conceptual change as measured by the pre to post distance between names for the same Fribble, averaged over two pair members (y-axis). More negative slopes (which we interpret as more or faster conceptual progress) are related to more conceptual change. The leftmost graph depicts the original exploratory results and the middle graph the current confirmatory results for conceptual change (within individuals). The smaller graph on the right depicts the current exploratory results with conceptual convergence as outcome measure, as measured by the pre to post change in distance between pair members' names for the same Fribble. Thus the power exponent of the curvilinear power function is related to communicative outcomes, indexed by both conceptual change and convergence.

[4] Random slopes were not included in any of our models given convergence issues.

Power coefficient and convergence

Exploratory results

A linear mixed-effects model with conceptual convergence as outcome measure, pair and Fribble as random intercepts, and average slope over all four effort measures as fixed effect, showed an effect of slope (β = -.20 ; SE = .053; *t*(641.4) = -3.78; *p* < .001), which indicates that the steeper (more negative) the slope and thus the more and/or faster the effort decreases over rounds, the more convergence takes place in Fribbles' conceptualizations between pair members (see Figure 4, right panel). One might say that convergence of conceptualizations is an even stronger operationalization of communicative effect than change of conceptualizations within individuals, since the change could occur in any direction. In the context of the present referential communication game, it is very beneficial to the goal of the game for pairs to establish "conceptual pacts" (Brennan & Clark, 1996). Here, we showed that conceptual progress (i.e., decrease of effort over the rounds/negative slopes in the interaction) is also related to the resulting conceptual convergence between communicators as measured by comparing pre- to post-interaction labeling of Fribbles.

General Discussion

The present study set out to investigate the pattern of multimodal reduction of *joint* communicative effort in an interactive referential game and its relation to communicative outcomes. We confirm our original findings (Bögels et al., 2023) that a negative power-law could describe the reduction of communicative effort over repeated references better than a linear description, which was also the case for external datasets (e.g., Hawkins et al., 2020). Crucially, this negative power law relationship is communicatively relevant - a more negative power exponent (i.e., a stronger/faster decrease of communicative effort and a high invested effort at beginning) relates to larger communicative outcomes. As such, we show that automatically trackable statistical markers of joint effort can be directly predictive of successful communication.

We establish strong evidence that communicative reductions in referential games are often not linear, even though they are generally modeled as such. Rather, communicative effort reductions fit well with a negative power law, where its power coefficient seems to reflect communicative work. Given that all modalities showed roughly the same pattern, the effort reduction seems to be a thoroughly multimodal phenomenon. Delving deeper into a similar phenomenon within psychology, a closest analogy might be with Thorndike's cat in a box 'escape curve' (Thorndike, 1898). Perhaps there are deep similarities to finding your way out of a maze and finding your way to a referent, which relate to offering proposals for potential solutions (variety), and selection based on optimal solutions that bring you closer to a goal. In the cat's situation, the more that is tried out, the more likely that one of those attempts turns out to be an optimal solution, bringing the cat closer to the goal. In communication, the more that is (collaboratively) proposed, the more likely it is that one of these proposals turns out to be a successful conceptual pact, which both participants understand and can remember. Of course one should keep in mind that, in communication (as opposed to a cat escape box), there are multiple possible solutions, the selection process is interactive, and the optimization process may continue over time.

So how can we make sense of reduction in language games like these? In simple reinforcement-based approaches (Erev & Roth, 1998) it is suggested that more informative elements are positively rewarded in interaction and then selected for re-expression (Beuls & Steels, 2013). What these models lack is an explanation of why longer descriptions are preferred at start, or why reductions do not occur when references are designed for outsiders (Hupet & Chantraine, 1992). It seems such minimal models lack a way that an agent reasons about the other agent. Hawkins, Franke and colleagues (2023) therefore propose a Bayesian agent-based model which sets up interacting agents as producing references based on reasoning about the other agent's lexicon while following Grice's efficiency maxim. In such simulations, agents provide longer descriptions since multiple references for an object decrease the ambiguity that individual elements/words would have on their own, given that at the beginning the director does not know the matchers lexicon (Hawkins, Franke, et al., 2023). After interaction, the director's priors about the matcher's lexicon are updated and some individual elements increase in their informativity relative to other elements. Since the agent is programmed to be maximally informative with minimal effort, at critical thresholds of increased differential informativity, shorter descriptions of maximally informing elements become preferred. Interesting in the Hawkins and colleagues simulations, the characteristic non-linearity of reductions seems to be replicated, though this may be due to the inbuilt logarithms used in their calculation of informativity.

In our terminology, borrowing from concepts in formal models in genetic algorithms (Adami, 2024), we conceptualize the process of translating variety of descriptions into shared conceptualizations through collaborative selection as communicative work: the degree to which you can transform raw energy (potential candidate descriptions) into effective energy that performs some function (effective references that lead to selection of the correct referent). Some pairs will not generate enough raw materials to work with (low variety that is unlikely to contain a high-fitness solution), some fail to select the effective solutions (low selection) thereby failing to reduce the variety, some will be able to find an optimum of both, which together reflects an ability to find a fit enough description that allows interlocutors to do the joint task of indexing objects based on short communications.

The current results provide a further handle for agent-based models to distinguish the different reasons for curvilinearity and how it relates to communicative success. Firstly, if a simulation model is able to account for a negative power law and crucially this can be related to communicative effects, it is more likely that it captures some of the dynamics observed in the wild. Note that generating a curvilinear relation is not enough and might be easy to reproduce: analytically we already know that if a model simply entails a constant proportional reduction over the rounds, this would approximate an exponential (closely following a logarithmic) relation and not a power relation. To get a functional understanding of the current communicative power law, agent-based models could be devised to ascertain how communicative success relates to investment and reduction of effort. Ideally, such models would make use of more general principles that might be applied to explain similar nonlinear phenomena in psychology and linguistics.

Our findings have been obtained in a specific communicative situation, a referential communication game with novel, difficult-to-describe objects. Our exploratory analyses of other datasets suggest that a similar curvilinear relation is found under different communicative situations (chat, audio, face-to-face) and referents (Fribbles, tangrams of different difficulties). We argue that everyday communication bears resemblances to such games in the sense that the meaning of any reference emerges from a collaborative process (Roberts & Bavelas, 1996; Stewart, 1996) and the amount of such 'negotiation' is likely to decrease over repetitions of the reference (within a conversation). Future research may further explore which other factors (e.g., interactivity, referent difficulty, communicative partner) may affect the exact shape of the reduction and its relation with communicative outcomes. For now, we can conclude that communicative effort reduction in referential communication games follows a particular pattern which is predictive of communicative outcomes. As such, we have found automatically trackable statistical markers of successful communication that can have important implications for understanding human communication from simple domain-general principles.

References

Adami, C. (2024). *The Evolution of Biological Information: How Evolution Creates Complexity, from Viruses to Brains*. Princeton University Press.

Alibali, M. W., Nathan, M. J., Wolfgram, M. S., Church, R. B., Jacobs, S. A., Johnson Martinez, C., & Knuth, E. J. (2014). How Teachers Link Ideas in Mathematics Instruction Using Speech and Gesture: A Corpus Analysis. *Cognition and Instruction*, *32*(1), 65–100. https://doi.org/10.1080/07370008.2013.858161

Arnold, J. E., Kaiser, E., Kahn, J. M., & Kim, L. K. (2013). Information structure: Linguistic, cognitive, and processing approaches. *WIREs Cognitive Science*, *4*(4), 403–413. https://doi.org/10.1002/wcs.1234

Bak, P. (1996). *How Nature Works*. Springer. https://doi.org/10.1007/978-1-4757-5426-1

Bangerter, A., Mayor, E., & Knutsen, D. (2020). Lexical entrainment without conceptual pacts? Revisiting the matching task. *Journal of Memory and Language*, *114*, 104129. https://doi.org/10.1016/j.jml.2020.104129

Barry, T. J., Griffith, J. W., De Rossi, S., & Hermans, D. (2014). Meet the Fribbles: Novel stimuli for use within behavioural research. *Frontiers in Psychology*, *5*.

https://doi.org/10.3389/fpsyg.2014.00103

Beuls, K., & Steels, L. (2013). Agent-Based Models of Strategies for the Emergence and Evolution of Grammatical Agreement. *PLOS ONE*, *8*(3), e58960. https://doi.org/10.1371/journal.pone.0058960

Boersma, P., & Weenink, D. (2012). *Praat: Doing phonetics by computer*. http://www.praat.org

Bögels, S., Li, T., Rasenberg, M., & Pouw, W. (2023). *Putting in the effort: Predicting communicative work based on joint communication effort and conceptual affordances (preliminary results and pre-registration)*. https://doi.org/10.17605/OSF.IO/ZHXB6

Branigan, H. P., Catchpole, C. M., & Pickering, M. J. (2011). What makes dialogues easy to understand? *Language and Cognitive Processes*, *26*(10), 1667–1686. https://doi.org/10.1080/01690965.2010.524765

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482–1493. https://doi.org/10.1037/0278-7393.22.6.1482

Clark, H. H., & Haviland, S. E. (1977). *Comprehension and the given-new contract Discourse Production and Comprehension (Vol. 1, pp. 1-40). Norwood*. NJ: Ablex Publishing Corporation.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

de Ruiter, J. P., Bangerter, A., & Dings, P. (2012). The Interplay Between Gesture and Speech in the Production of Referring Expressions: Investigating the Tradeoff Hypothesis. *Topics in Cognitive Science*, *4*(2), 232–248. https://doi.org/10.1111/j.1756-8765.2012.01183.x

Ebbinghaus, H. (1885). *Memory*. Рипол Классик.

Eijk, L., Rasenberg, M., Arnese, F., Blokpoel, M., Dingemanse, M., Doeller, C. F., Ernestus, M., Holler, J., Milivojevic, B., Özyürek, A., Pouw, W., van Rooij, I., Schriefers, H., Toni, I., Trujillo, J., & Bögels, S. (2022). The CABB dataset: A multimodal corpus of communicative interactions for behavioural and neural analyses. *NeuroImage*, *264*, 119734. https://doi.org/10.1016/j.neuroimage.2022.119734

Erev, I., & Roth, A. E. (1998). Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *The American Economic Review*, *88*(4), 848–881.

Fowler, C. A. (1988). Differential Shortening of Repeated Content Words Produced in Various Communicative Contexts. *Language and Speech*, *31*(4), 307–319. https://doi.org/10.1177/002383098803100401

Fowler, C. A., & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, *26*(5), 489–504. https://doi.org/10.1016/0749-596X(87)90136-7

Galati, A., & Brennan, S. E. (2014). Speakers adapt gestures to addressees' knowledge: Implications for models of co-speech gesture. *Language, Cognition and Neuroscience*, *29*(4), 435–451. https://doi.org/10.1080/01690965.2013.796397

Gerwing, J., & Bavelas, J. (2004). Linguistic influences on gesture's form. *Gesture*, *4*(2), 157–195. https://doi.org/10.1075/gest.4.2.04ger

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R.

(2019). How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, *23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003

Gijssels, T., Casasanto, L. S., Jasmin, K., Hagoort, P., & Casasanto, D. (2016). Speech Accommodation Without Priming: The Case of Pitch. *Discourse Processes*, *53*(4), 233–251. https://doi.org/10.1080/0163853X.2015.1023965

Gilden, D. L. (2001). Cognitive emissions of 1/f noise. *Psychological Review*, *108*(1), 33–56. https://doi.org/10.1037/0033-295X.108.1.33

Givón, T. (1983). *Topic Continuity in Discourse*. 1–498.

Grice, H. P. (1975). *Logic and Conversation*. https://doi.org/10.1163/9789004368811_003

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020). Characterizing the Dynamics of Learning in Repeated Reference Games. *Cognitive Science*, *44*(6), e12845. https://doi.org/10.1111/cogs.12845

Hawkins, R. D., Franke, M., Frank, M. C., Goldberg, A. E., Smith, K., Griffiths, T. L., & Goodman, N. D. (2023). From partners to populations: A hierarchical Bayesian account of coordination and convention. *Psychological Review*, *130*(4), 977–1016. https://doi.org/10.1037/rev0000348

Hawkins, R. D., Sano, M., Goodman, N. D., & Fan, J. E. (2023). Visual resemblance and interaction history jointly constrain pictorial meaning. *Nature Communications*, *14*(1), 2199. https://doi.org/10.1038/s41467-023-37737-w

Hincks, R. (2005). Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism. *System*, *33*(4), 575–591. https://doi.org/10.1016/j.system.2005.04.002

Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E., & Swerts, M. (2015). Reduction in gesture during the production of repeated references. *Journal of Memory and Language*, *79–80*, 1–17. https://doi.org/10.1016/j.jml.2014.10.004

Holler, J., Bavelas, J., Woods, J., Geiger, M., & Simons, L. (2022). Given-New Effects on the Duration of Gestures and of Words in Face-to-Face Dialogue. *Discourse Processes*, *59*(8), 619–645. https://doi.org/10.1080/0163853X.2022.2107859

Holler, J., Tutton, M., & Wilkin, K. (2011). *Co-speech gestures in the process of meaning coordination*. 2nd GESPIN - Gesture & Speech in Interaction. https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_1105560

Holler, J., & Wilkin, K. (2011). Co-Speech Gesture Mimicry in the Process of Collaborative Referring During Face-to-Face Dialogue. *Journal of Nonverbal Behavior*, *35*(2), 133–153. https://doi.org/10.1007/s10919-011-0105-6

Hupet, M., & Chantraine, Y. (1992). Changes in repeated references: Collaboration or repetition effects? *Journal of Psycholinguistic Research*, *21*(6), 485–496.

Jacobs, N., & Garnham, A. (2007). The role of conversational hand gestures in a narrative task. *Journal of Memory and Language*, *56*(2), 291–303. https://doi.org/10.1016/j.jml.2006.07.011

Kauffman, S. A. (2019). *A World Beyond Physics: The Emergence and Evolution of Life*. Oxford University Press.

Levshina, N., & Moran, S. (2021). Efficiency in human languages: Corpus evidence for universal principles. *Linguistics Vanguard*, *7*(s3). https://doi.org/10.1515/lingvan-2020-0081

Levy, E. T., & McNeill, D. (1992). Speech, gesture, and discourse. *Discourse Processes*, *15*(3), 277–301. https://doi.org/10.1080/01638539209544813

MacIntyre, A. D., Cai, C. Q., & Scott, S. K. (2022). Pushing the envelope: Evaluating speech rhythm with different envelope extraction techniques. *The Journal of the Acoustical Society of America*, *151*(3), 2002–2026. https://doi.org/10.1121/10.0009844

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Marcoux, K., & Ernestus, M. (2019). PITCH IN NATIVE AND NON-NATIVE LOMBARD SPEECH. *19th International Congress of Phonetic Sciences (ICPhS 2019)*, 2605–2609.

Marslen-Wilson, W., Levy, E., & Tyler, L. K. (1982). Producing interpretable discourse: The establishment and maintenance of reference. *Speech, Place, and Action*, 339–378.

Motamedi, Y., Schouwstra, M., Smith, K., Culbertson, J., & Kirby, S. (2019). Evolving artificial sign languages in the lab: From improvised gesture to systematic sign. *Cognition*, *192*, 103964. https://doi.org/10.1016/j.cognition.2019.05.001

Pouw, W., Dingemanse, M., Motamedi, Y., & Özyürek, A. (2021). A Systematic Investigation of Gesture Kinematics in Evolving Manual Languages in the Lab. *Cognitive Science*, *45*(7), e13014. https://doi.org/10.1111/cogs.13014

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org

Rasenberg, M., Özyürek, A., Bögels, S., & Dingemanse, M. (2022). The Primacy of Multimodal Alignment in Converging on Shared Symbols for Novel Referents. *Discourse Processes*, *59*(3), 209–236. https://doi.org/10.1080/0163853X.2021.1992235

Rasenberg, M., Pouw, W., Özyürek, A., & Dingemanse, M. (2022). The multimodal nature of communicative efficiency in social interaction. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-22883-w

Roberts, G. L., & Bavelas, J. B. (1996). The communicative dictionary: A collaborative theory of meaning. In J. Stewart (Ed.), *Beyond the symbol model* (pp. 135–160).

Stevens, S. S. (1961). To honor fechner and repeal his law: A power function, not a log function, describes the operating charactersitic of a sensory system. *Science*, *133*(3446), 80–86. https://doi.org/10.1126/science.133.3446.80

Stewart, J. R. (1996). *Beyond the Symbol Model: Reflections on the Representational Nature of Language*. SUNY Press.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4), i–109. https://doi.org/10.1037/h0092987

Vajrabhaya, P., & Pederson, E. (2018). Teasing apart listener-sensitivity: The role of interaction. *Gesture*, *17*(1), 65–97. https://doi.org/10.1075/gest.00011.vaj

Viering, T., & Loog, M. (2023). The Shape of Learning Curves: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(6), 7799–7819. https://doi.org/10.1109/TPAMI.2022.3220744

Weber, E. H. (1978). *E. H. Weber: The Sense of Touch*. Academic Press for Experimental Psychology Society.

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., … Li, J. (2017). CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 1–19. https://doi.org/10.18653/v1/K17-3001

Zipf, G. K. (1949). *Human behavior and the principle of least effort* (pp. xi, 573). Addison-Wesley Press.

Supplementary Materials

S1. Extended Methods

Participants/Pairs

The dataset (Eijk et al., 2022) consists of 42 pairs of participants for which the individual pre and post tasks can be used, participants gave consent to share audio-video recordings of the interactions with external researchers, and transcriptions of the interaction are available (including annotations of trials etc.), which are all necessary for our analyses. These pairs consist of 17 all-female pairs, 5 all-male pairs, and 20 mixed pairs. The participants were 22.4 years old on average (SD = 3.02, range = 18-33).

Overview of the paradigm

Two participants, who did not know each other beforehand, came to the lab. After giving informed consent, they performed several individual tasks, then interacted together, and then again performed individual tasks. The individual tasks consisted of a Naming and a Features tasks (see below), which were performed alternately per Fribble. In addition, participants performed a task in the MRI scanner in which they viewed the Fribbles one by one on the screen, each 12 times in total (lasting about 30 minutes) both right before and right after the interaction. The paradigm also included a short individual phonological pretest just before the interaction as well as another fMRI session at the end in which participants viewed eight short animated movies (35 minutes). Finally, participants filled out a questionnaire, about strategies used in different parts of the experiment, impressions of the pair member, and questions about the participants themselves (see Table S3 in Supplementary Materials of Eijk et al. (2022) for a complete list of the questions).

Naming task

The Naming task consisted of 16 trials, randomized per participant. In each trial, participants saw all 16 Fribbles on the screen (in a randomized order per participant), with a red square around the target Fribble. They were instructed to name the target Fribble by typing in aname on the keyboard, using one to three words, in such a way that their partner would be able to correctly identify the Fribble amongst the others.

Features task

The Features task also consisted of 16 trials (interspersed with the Naming task). In each trial, the target Fribble was presented in the left corner of the screen and participants were asked to rate the target Fribble on 29 different features displayed on the screen. Examples of rated features are: “To what extent do you see this picture as rounded/symmetrical/easily audible/human?” (see Table S1 in the Supplementary Materials of Eijk et al. (2022) for the complete list of features). Underneath each feature, a horizontal bar was visible with a small vertical bar on top that could be moved using the mouse. Participants were asked to change the position of the vertical bar to indicate to what extent the feature matched their own view of the Fribble. In case they thought the feature was not applicable, they were asked to put the vertical bar to the very left. The position of the bar was transformed into a score between 0 (fully left) and 100 (fully right).

Interaction: Referential task

The interaction between the two participants consisted of 16 trials (16 target Fribbles) times 6 rounds, so 96 trials in total. Participants first performed the Referential task, immediately followed by the Localisation task (not relevant for the present study) in each trial. Participants saw all 16 Fribbles on their screen at the same time, randomly distributed over the screen (which was necessary for the Localisation task, see Eijk et al., Figure 4). Eight Fribbles were placed at the same position on both screens and the other eight at different positions. The arrangement of Fribbles on the screen changed after every round. In each trial, one participant (the Director) saw a red square around the target Fribble and was asked to describe that Fribble to the other participant (the Matcher). They were instructed that they could communicate in any way that they wanted and that the Matcher was allowed to ask questions. Once the Matcher had identified the Fribble on their screen, they were asked to name the label (letter or number) corresponding to that Fribble out loud and press a button on a button box to continue. In the Localization task (which was not part of the smaller dataset used for our original exploratory analyses), participants were asked to determine together whether the target Fribble was located at the same or a different position on both their screens. The interaction lasted 52.19 minutes on average (range = 35.20-77.48 minutes).

Preprocessing and operationalizations

Below, we explain in detail our operationalizations for the communicative effort (four different types of effort measures), communicative effects (i.e., average pre-post change in naming and pre-post convergence in naming between pair members), and conceptual affordances (i.e., centrality of the referents). Preprocessing and data analysis was done mainly in R version 4.2.2 (R Core Team, 2022). Any other used software is mentioned where appropriate.

Measures of communicative effort

To quantify the pairs' efforts for each trial, we defined four basic and crude effort measures in our analysis that could be automatically extracted, described in turn below. All measures, except for transitions (interactional effort), were measured per participant and then summed over the two participants of a pair for most analyses reported below.

Speech effort. To operationalize speech effort, we extracted the speech amplitude envelope from the audio recordings of the interactions for each participant (Pouw & Trujillo, 2021), with a smoothing Hanning filter of 5Hz and a resampling rate of 100Hz. We then counted the number of envelope peaks per trial as a proxy for the number of syllables (see e.g., MacIntyre et al., 2022), with a peak height threshold of 0.37 (set by M-(2*SD) envelope height over all participants).

Gesture Effort. For gestural effort, we used a measure based on participants' movements, which have previously been confirmed to correlate well with the amount of annotated iconic gestures (Rasenberg, Özyürek, et al., 2022), though they include any type of manual movement during the conversation. Specifically, we computed the average speed of participants' left and right hand tips over time using coordinates as recorded using a Microsoft Kinect (V2) system, then counted the number of movement speed peaks per trial, with a peak distance threshold of 200 milliseconds and a height threshold of 15 cm per seconds (Pouw et al., 2021; Rasenberg, Pouw, et al., 2022).

Interactional Effort. For interactional effort, we counted the number of speaker turn transitions for each trial using time-stamped speech transcripts. When a turn fully overlapped with a longer turn of the other participant (including backchannels, like "uh-huh" or "yeah"), this was counted as two transitions.

Prosodic Effort. For prosodic effort, we extracted the F0 of each individual's audio recording using Praat (Boersma & Weenink, 2012), with a sample rate of 100Hz. Pitch range thresholds were 50–250 Hz for male speakers and 80–350 Hz for female speakers (Gijssels et al., 2016). Then the raw values were cleaned to avoid pitch doubling and halving, following a Python script by Marcoux and Ernestus (2019). We computed PVQ (Pitch Variation Quotient, pitch SD/M; Hincks, 2005) as a measure for the pitch variation of the participant in each trial.

Measures of communicative effect

We used the names from the Naming task administered before and after the interaction to estimate the change in the participants' conceptualizations of the Fribbles and the convergence between two pair members over the interaction. To measure semantic dsitances between the names, we used a word2vec distance measure as implemented in SNAUT (Mandera et al., 2017). The names were first corrected by removing characters (indicated here between <>, e.g. <'>, <">, <()>, <&>, <+>, <.>, <;>), converting characters <-> and </> into a space, <=> into the word <is>, correcting obvious spelling errors, changing uppercase into lowercase characters, and changing numbers into the corresponding number words (Eijk et al., 2022). Using SNAUT, the words were then checked against the corpus used for word2vec calculations (see below). Words missing from the corpus were corrected when mis-spelled, and missing compounds were divided into two (or more) words. We used the pre-trained corpus NLPL Dutch CoNLL17 (2.610.658 items, 100 dimensions; Zeman et al., 2017) and computed the cosine distance between the names for the same Fribble given by the same participant before and after the interaction using SNAUT (i.e., change). To compute average change, we then averaged the distances over the two members of a pair per Fribble. We also computed the cosine distance between the names for the same Fribble given by the two participants of a pair, both before and after the interaction. To compute the convergence between pair-members we subtracted the post distance from the pre distance (such that larger numbers correspond to relatively smaller distances post and thus more convergence).

Measures of conceptual affordances

As a measure of the conceptual affordances of a Fribble for a specific pair, that is, how difficult it would be for them to change each other's conceptualizations for that Fribble, we used the relative similarity of the Fribble to the other Fribbles, considering that the more similar a Fribble is to the other Fribbles, the more difficult it is to refer to it and change its conceptualization. As a purely visual measure (irrespective of the participants) we used pixel-wise similarity scores (structural similarity) between pairs of Fribbles. We created a visual distance score for each Fribble by averaging over the distances to the other 15 Fribbles.

To create a pair-dependent measure of conceptual affordances, we used the semantic distance (word2vec cosine distance, see above) between pre-interaction names to measure the semantic distance from one Fribble name to another for each pair. These distances were calculated between the 32 pre-interaction names (16 Fribbles by 2 pair members) from each pair.

To create a pre-distance score for each Fribble, we averaged over the 31 distances to all other names and then averaged these scores over the two members of the pair. We reason that the more similar a Fribble name, given before the interaction, is to the names given to all other Fribbles by this specific pair, the more difficult it will be for this pair to refer to that Fribble and thus change each other's conceptualization for this Fribble.

S2. Current non-replicated confirmatory hypotheses, results and discussion

Hypotheses

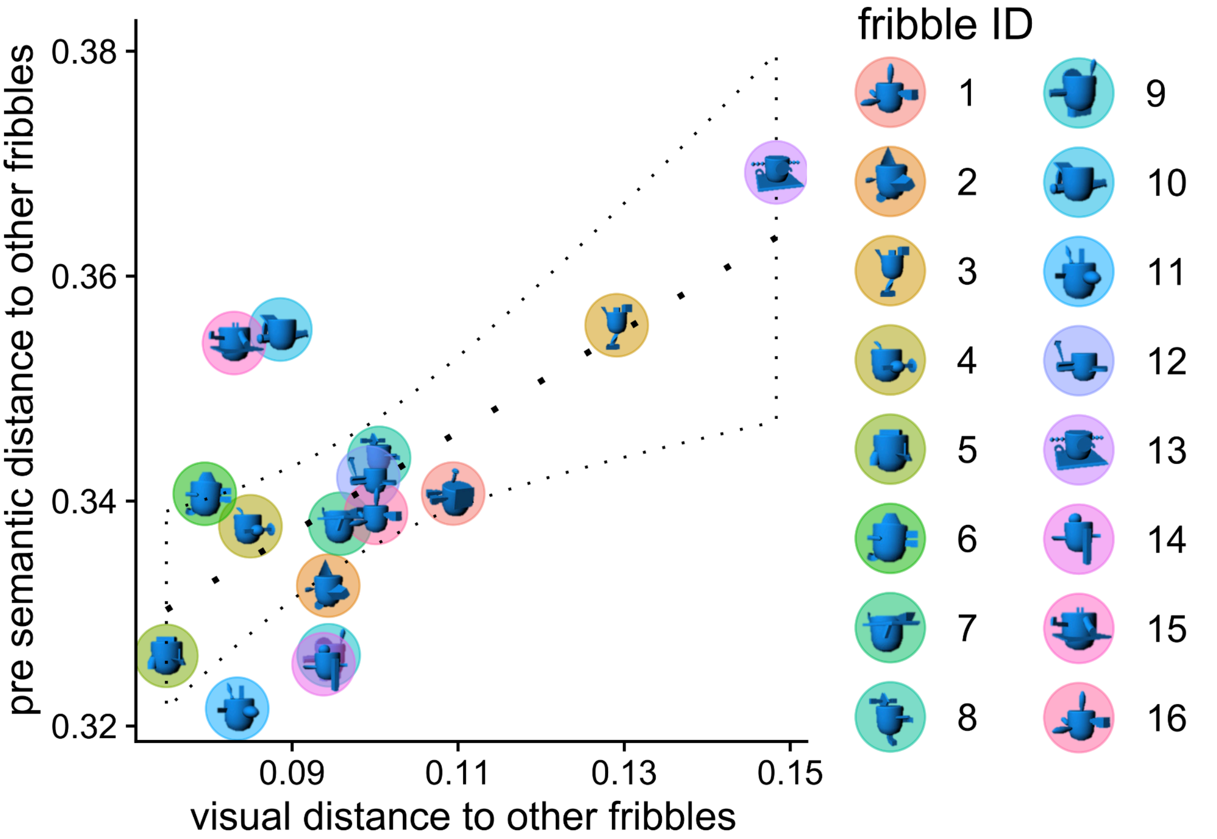
The hypotheses below appeared as hypotheses 2 and 3, respectively, in our preregistration {Updating}eff.

3. Fribbles that are visually more different from other Fribbles also elicit names that are more semantically different from the other Fribbles.
4. Fribbles' conceptual affordances moderate the effect of average effort on conceptual change such that for easy Fribbles (which are conceptualized differently from other Fibbles) effort is positively related to change, whereas for difficult Fribbles (which are conceptualized similarly to other Fribbles) this relation may be non-existent or negative.

Results and Discussion

Regarding Hypothesis 3, our confirmatory analyses did not show a correlation between the pre-distance score and visual distance ($r = -.23$, $p > .05$), a correlation that was significant in our pre-registration. A Mantel test comparing the pre–naming scores (averaged over both participants) and visual distance matrices directly also did not show a significant similarity (z = 4.64; p > .05). This means we cannot confirm Hypothesis 3. Looking at this correlation in the original exploratory dataset (see Figure S1), it appears that it is mainly driven by only two or three of the Fribbles that are visually quite distinct from the others, which may have rendered the finding relatively unstable. In our exploratory dataset, this distinction was apparently picked up by many participants and reflected in their naming of these Fribbles when they first saw them (before the interaction), while the names in the larger confirmatory dataset appeared less affected by visual characteristics of the Fribbles. A speculative explanation may be that participants in the confirmatory dataset performed the Naming task interspersed with the Features task in which they were asked to judge each Fribble on 29 features ranging from visual to more abstract. This may have prompted them to think about the Fribbles more abstractly.

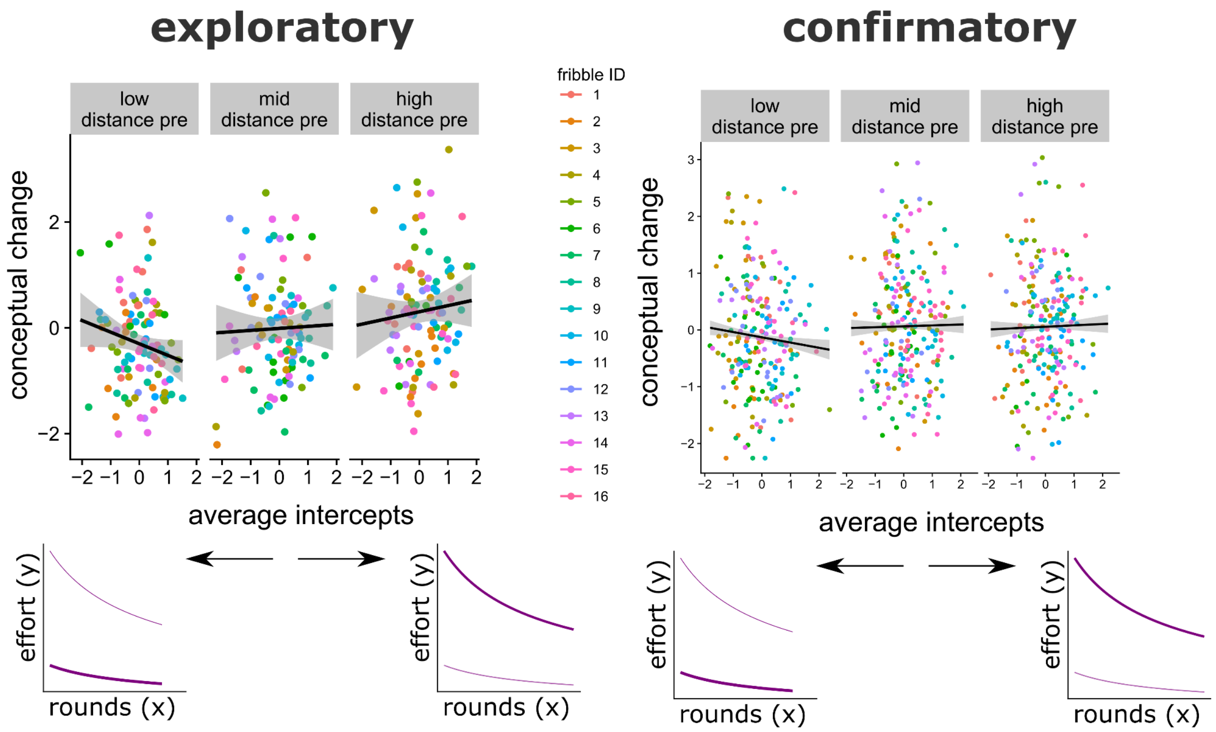
Figure S1. Average visual distance and pre semantic distances



Note. Results of the original exploratory analyses regarding Hypothesis 3 (which could not be replicated by our current confirmatory analyses). The graph plots the semantic distances between each Fribble and all other Fribbles, averaged over all pairs (y-axis) against the visual distances between each Fribble image and all other Fribble images (averaged) as calculated using structural similarity (x-axis). Fribble identities are also color coded. The trend indicates that in general, Fribbles that have very different visual features from the other Fribbles, also are conceptualized about more differently relative to other Fribbles before the interaction.

Regarding Hypothesis 4, a linear mixed-effects model with change as outcome measure, pair, and Fribble as random intercepts, and average intercept, pre-naming distance, and their interaction as fixed effects did not show a significant effect for both of the fixed effects (average intercepts: β =-0.41; SE =0.31; t(658.89) = -1.33; p = .18, pre-naming distances: β = -0.67; SE = 0.76; t(303.04) = -0.87; p = .38). This indicates that neither the amount of communicative effort, nor conceptual affordances of the Fribble relate to conceptual change. There was also no interaction (β =0.86; SE =0.79; t(662.22) =1.08; p = .28). As can be seen in Figure S2, the trends that were significant in the exploratory dataset appear similar in the large dataset, but they do not reach significance, so Hypothesis 4 cannot be confirmed. It is possible that the way we operationalized the conceptual affordances of the Fribbles may not have captured (all) the relevant aspects of what makes some objects easier to describe than others. In the confirmatory dataset, the pre-naming distances were not related to visual distances. It is possible that participants' conceptualizations of the Fribbles were thus less related to the visual characteristic of the Fribbles and more idiosyncratic. This may also have made them more different between participants. Within a specific pair, it may be more difficult to describe a Fribble to your partner when their initial conceptualization is very different from yours. This aspect was not incorporated in our operationalization of the Fribbles' conceptual affordances, for example. Future work may take into account such interactional and other aspects that may affect conceptual affordances.

Figure S2. Average intercept (effort), semantic distance pre (conceptual affordances), and conceptual change



Note. The graph shows the relation between average intercepts (effort low to high, x-axis) and the conceptual change as measured by the averaged pre to post distance between names for the same Fribble, averaged over two pair members. For visualization purposes only, the data points were split into three equal parts based on their corresponding pre semantic distances (the three panels). The right panel shows our original exploratory results. There, we found that for low pre-distances (Fribbles that are difficult to differentiate conceptually), higher intercepts (i.e., more effort) are related to less conceptual change (left small graph). For higher pre-distances (Fribbles that are more easily conceptually differentiable), higher intercepts (more effort) are related to more conceptual change (middle and right small graphs). These results could not be replicated in our current confirmatory analyses (right panel), although the pattern of results appears similar.