Coordination, rather than pragmatics, shapes colexification when the pressure for efficiency is low.

Alexey Koshevoy[1,2], Isabelle Dautriche[1], Olivier Morin[2] and Kenny Smith[3]
[1]Centre de Recherche en Psychologie et Neurosciences, Aix-Marseille Université, Marseille, France
[2]Institut Jean Nicod, Département d'études cognitives, ENS-PSL, Paris, France
[3]School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, Scotland

Abstract

We investigate the phenomenon of colexification, where a single wordform is associated with multiple meanings. Previous research on colexification has primarily focused on empirical studies of different properties of the meanings that determine colexification, such as semantic similarity or meaning frequency. Meanwhile, little attention was paid to the wordforms' properties, despite being the original approach advocated by Zipf. Our preregistered study examines whether word length influences word choice for colexification using a novel dyadic communication game (N = 64) and a computational model grounded in the Rational Speech Act (RSA) framework. Contrary to initial predictions, participants did not exhibit a strong preference for efficient colexification (namely colexifying multiple concepts using short words, when long alternatives are available). The results align more closely with a simpler coordination model, where dyads align on a functioning lexical convention with relatively little influence from the efficiency of that convention. Our study highlights the possibility that colexification choices are strongly determined by the pressure for coordination, with weaker influences from semantic similarity or meaning frequency. This is most likely explained by weak pressure for efficiency in our experimental design.

Keywords: colexification; Rational Speech Acts Framework; dyadic communication game; computational modelling

Introduction

The mapping between wordforms and meanings may differ substantially from one language to another. For instance, in English the verb *go* means both "go by foot" and "go by vehicle", while in German the former is expressed by the verb *gehen*, and the latter by the verb *fahren*. The phenomenon is denoted by terms such as ambiguity, polysemy, or colexification amongst many others (Haspelmath, 2023). In this paper, we will refer to this phenomenon using the term *colexification*, which denotes a situation in which a single word-form corresponds to multiple meanings, like the word *go* in English (François, 2008). What factors motivate particular colexifications? Brochhagen and Boleda (2022) suggested that meanings are colexified if they strike a balance between discriminability and economy, avoiding both excessive similarity and dissimilarity. Furthermore, Karjus, Blythe, Kirby, Wang, and Smith (2021) provided experimental evidence that, while semantic similarity between meanings leads participants to preferentially colexify those similar meanings, communicative need, i.e. the importance of conveying distinctions between particular meanings, fosters the inverse process[1]. However, the properties of the *wordforms* used for colexification have been explored to a much lesser extent. For instance, why does English use the wordform *go* to colexify "go by foot" and "go by vehicle" instead of *drive*, which corresponds to only one of these meanings ("go by vehicle")? The original explanation to this was provided in Zipf (1949). In his work, G.K. Zipf introduced the idea of a trade-off between speaker's and hearer's economies that shapes natural languages. On the one hand, speakers aim to convey their intended meaning efficiently, by expending the least amount of effort possible. On the other hand, hearers aim to infer the intended meaning accurately. These goals can sometimes be in conflict, as speakers may want to use less specific or ambiguous language to convey their intended meaning, while hearers may need to expend more effort to infer the intended meaning from the available linguistic cues. To balance between these two opposing forces, natural languages may display some degree of ambiguity. Moreover, Zipf predicted that these systems would favor shorter words for ambiguous terms; as ambiguous terms cover multiple meanings they will be used more frequently, favouring short, less effortful wordforms (Zipf, 1945). Zipf's predictions about the relation between length and number of meanings were empirically tested by Piantadosi, Tily, and Gibson (2012). However, this analysis does not offer a mechanistic explanation for the observed preference for efficiency when colexifying multiple meanings. In other words, there is to date no account explaining this phenomenon from the perspective of the process of actual communication between individuals, such as the one presnted in Kanwal, Smith, Culbertson, and Kirby (2017) for the Zipf's law of Abbreviation.

One possibility for a mechanistic explanation is a demonstration that principles posited by Zipf can be derived using the principles of pragmatic reasoning. This idea has a long history; for instance, Horn (1984, 1993) proposed that Zipf's conflicting forces (speaker's and hearer's economies) can be linked to Gricean maxims (Grice, 1975). He suggested equating Zipf's opposing forces with the Q principle (maximize informativeness, or hearer's economy) and R principle (minimize effort, or speaker's economy). The goal of our study is, therefore, to provide experimental evidence for the equivalence of Zipfian trade-offs and pragmatic reasoning processes. More specifically, we want to explore the relation between wordform length and the probability of using a word to colexify multiple meanings.

We designed a dyadic communication game where two

[1]I.e. use of more specified wordforms for each individual meaning (like in the German example)

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

participants are asked to communicate about three meanings (geometric shapes) using only two words. Presenting only two words for three meanings requires the participants to colexify two of the three meanings. We introduce context (as colour) which makes it possible for the participants to disambiguate between two out of three meanings even if they are labeled by the same word, favouring colexification of those two meanings. Lastly, we used long and short words (with longer words taking more time to send, in addition to being longer) to look into efficiency. Crucially, and contrary to similar experimental designs (Kanwal et al., 2017), participants cannot use meaning frequency to converge on particular word use pattern, like using short words for more frequent meanings. We also control for meaning similarity by using diverse sets of geometric objects as meanings. This experimental setup allows us to test whether participants prefer using short words to colexify multiple meanings, while controlling for meaning frequency and meaning similarity, which are known to affect colexification (Karjus et al., 2021).

To assess whether the expected preference to colexify using short wordforms can be derived from pragmatic principles, we propose a computational model based on the Rational Speech Act (RSA) framework (Frank & Goodman, 2012) that makes quantitative predictions about the behaviour of the participants in this experiment. RSA is a computational framework of language use that combines ideas from game theory and Bayesian cognitive science to formalize pragmatic reasoning. RSA reflects both the Quantity maxim and the maxim of Manner (Degen, 2023), making it a suitable theoretical foundation for equating Zipfian principles with principles of pragmatic reasoning. Recently, Peloquin, Goodman, and Frank (2020) have shown that pragmatic agents do indeed favour ambiguous mappings between words and meanings when context is informative about meaning, as predicted by G.K. Zipf and (Piantadosi et al., 2012), and that those systems are efficient. However, they did not explore the notion of the trade-off in greater detail, and it is also unclear whether the observed behaviour could be reproduced in human participants. We hypothesize that human participants prefer to use shorter words for colexifying multiple meanings, because it is more efficient. Additionally, we predict that participants' behavior aligns with that of pragmatic agents, who consider both the cost of messages and their informativeness.

Experiment

We used a dyadic communication game in a form of a web app[2] to test our hypotheses. This approach was previously used in several studies that examined different aspects of efficiency in the lexicon (Kanwal et al., 2017; Silvey, Kirby, & Smith, 2019; Karjus et al., 2021; Morin, Müller, Morisseau, & Winters, 2022), and more broadly for the study of language evolution in general (Scott-Phillips & Kirby, 2010; Galantucci, Garrod, & Roberts, 2012). The experimental methods and the analyses were preregistered and the preregistration, data and code are available `https://osf.io/a37vz/?view_only=0469f5b304f7409ab038a83aa9d7a256`.

[2]The app was implemented in Python and JavaScript using the Flask/Socket.io framework.

Participants

64 participants (32 dyads) were recruited via the Prolific platform. All of the participants were native English speakers. We excluded 2 dyads (4 participants) that did not use the context/color to inform their responses, see procedure below. Participants were paid a base rate of £4.20. They were told they could get an additional bonus of £4.20 if they were the fastest time to complete the task (note that all participants received the bonus). In addition, they received £0.10 for each correct response. This system of bonus payments was intended to ensure that participants were under pressures to be both accurate in communication, while also completing the task as fast as possible, i.e. to communicate efficiently.

Procedure

Participants were asked to encode and decode messages about different geometric shapes, playing the roles of senders and receivers respectively. The experimental pipeline is displayed on Figure 2. Senders were provided with a stimulus (a colored shape; e.g., a red circle), and were asked to choose one of the two words (one short, e.g., "nais", and one long, e.g., "uditslev") to describe it to the receiver. The difference between the two words is that one takes more time to be sent than the other, and this is reflected in the amount of time the sender needs to press on a virtual button (1 second for a short label, 4 seconds for a long label). The total time that the sender is taking to communicate is indicated on the top of the screen. Once the word is sent to the receiver, the receiver is presented with the color of the stimulus (e.g., red) and the word chosen by the sender (e.g., 'nais'). Then, the receiver's task was to choose the shape of the stimulus that was communicated by the sender given the information about the color. Both senders and receivers could use the "hint" button, which will display all possible stimuli (color-shape combinations) to make the guessing easier (the display is similar to Figure 1). After each round, both the sender and receiver were given feedback about whether the receiver's guess was correct or not, and they exchanged roles. A similar methodology was used in Kanwal et al. (2017) to show how Zipf's Law of Abbreviation (shorter elements becoming more frequent in a code) arises during communication, by incentivising the participants to be accurate while sending shorter messages. Our experiment consisted of 42 rounds, with each player alternating between the roles of sender and the receiver, thus playing for 21 rounds in each role.

Stimuli

Shapes The shapes are our proxy for meanings, and colors (which are provided to the receiver in additional to the sender's selected signal) are intended as a proxy for context that can be used to disambiguate between intended meanings. Figure 1 shows one example of a stimuli set for shapes. The

4 stimuli in a set includes 3 distinct shapes (e.g., pentagon, square and circle) that can have 2 distinct colors (e.g. purple or green), such that one of the shapes (e.g., the pentagon) occurs in two colors and the other two shapes only occur in one colour (squares are only ever green, circles are only ever purple). In the case of this set, receivers could use the information about color to effectively disambiguate between the square and the circle, but not between pentagons and any other shape, since the pentagon appears in both colors. Perfect communication is therefore possible in this scenario, but only by using context (colour) to disambiguate where possible and when there is one word that is used to refer to both square and circle (i.e. they are colexified), and the other word is used exclusively to refer to the pentagon. In this configuration, we will refer to the pentagon as the stimuli where **color is uninformative** about shape (**color-uninformative shape**), and the circle and the square as the stimuli where **color is informative** about shape (**color-informative shape**). Let's consider an example; in the case of the shape set on Figure 1, if a receiver is provided with the information that the color of the shape is purple, the only two possibilities for a response are circle and pentagon. If the two participants have previously managed to form a convention in which $word_1$ corresponds to color-informative shapes (square and circle), while $word_2$ corresponds to color-uninformative shape (pentagon), then the receiver can correctly choose the shape that was encoded by the sender by disambiguating between square and circle based on provided colour if the sender sent $word_1$, or choosing the pentagon if the sender sent $word_2$.

Each of the three shapes has a $1/3$ probability of being shown to the sender, meaning that all shapes are equiprobable (in the experiment, each shape appeared exactly $42/3 = 14$ times, with no more than two repetitions in a row). Although the color-uninformative stimuli can appear in two colors (purple and green) that each appears with a frequency of $1/6$, the corresponding shape still has the frequency of $1/3$. To control for the possible interference effects of shapes on the participant word choice (e.g. through iconicity) (Lewis & Frank, 2016; Xu, Duong, Malt, Jiang, & Srinivasan, 2020; Lewis & Frank, 2016), we generated 9 distinct sets of stimuli.

Words Participants were asked to choose between two pseudowords to communicate about the shapes to their partner. We obtained a list of 1.000 pseudowords of 4 and 8 letters from Rastle, Harrington, and Coltheart (2002). This list contains pseudowords generated in according to English phonotactics. We then filtered it to obtain a list of 10 randomly selected short and long word pairs where the edit distance is the highest, i.e. equal to 8. Each dyad received a randomly sampled pair consisting of a long and a short word. We used short words such as "cauv", "tarb" and "ciff", and long words such as "shoughse", "ghleente" and "ghleuche".

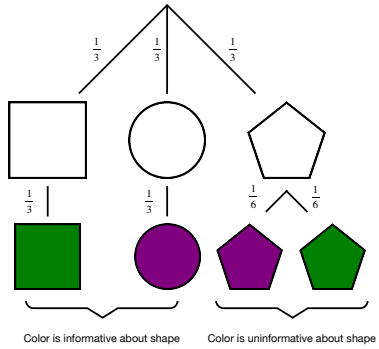

Figure 1: An example of a possible stimuli set in the experiment. The first row represents the 3 stimuli that are available to receivers, and the second row represents the 4 stimuli available to senders. Fractions indicate frequency of appearance to receivers and senders, respectively.

Measures

Communicative success To measure whether were communicatively successful[3], we computed the proportion of correct guesses after the burn-in period of 10 rounds. We set the color-informed accuracy baseline to be at 0.5, this baseline corresponds to randomly choosing one of the shapes by simply relying on color. It can be derived as follows: $\text{P(correct)} = \text{P(correct}|\text{color 1)} \times \text{P(color 1)} + \text{P(correct}|\text{color 2)} \times \text{P(color 2)} = \frac{1}{2} \times \frac{1}{2} \times 2 = \frac{1}{2}$.

Lexicon choice In this experiment, we were primarily interested in whether participants used the short or the long word with color-informative or color-uninformative stimuli. We measured the proportion of times participants used each word (long or short) for each stimulus type (color-informative or not), out of all the instances of each stimuli type. Therefore, each dyad got 2 scores; (a) the proportion of times the short word referred to color-informative stimuli, (b) the proportion of times the long word referred to one of the color-uninformative stimuli (both in the range from 0 to 1). The word which is used with the color-informative stimuli determines the preferences of participants with regards to colexification, since there are always two color-informative shape in the task, while there is only one color-uninformative shape. These proportions can be used to analyse the lexicons that the participants end up using. There are 4 possibilities: "efficient" mappings (short word for color-informative shape, long for color-uninformative), non-efficient mappings (long for color-informative, short for color-uninformative), "only short" (using only short words), and "only long" (using only long words). Only the efficient and non-efficient mappings can lead to communicative success; since there is only 1 short word, the "only short" lexicon reflects a desire to reduce average effort at the cost of sacrificing accuracy on trials where context does not disambiguate, whereas "only long" repre-

[3]I.e., created an accurate convention, instead of simply relying on color cues and sending random messages.

sents an anti-efficient lexicon with both increased effort and reduced communication success. These four lexicons can be represented by mapping each dyad on a 1 by 1 square, with proportion of color-uninformative shapes referred by long word on the x-axis, and proportion of color-informative shapes referred to using a short word, as shown on Figure 3 (panel A).

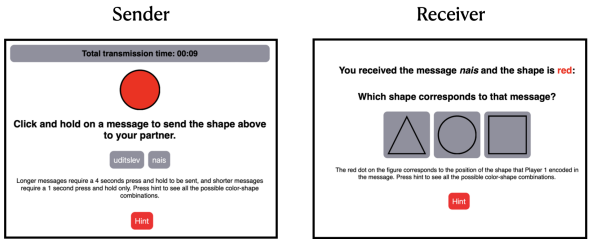

Figure 2: The sender view on a single communicative round (left) and the receiver view once the sender has selected and sent a signal (right).

Models

We constructed three different models that are used to predict the behaviour of the participants considering different hypotheses about the pressures shaping their behaviour. The pragmatic model considers speaker's economy trading off with hearer's economy (Quantity maxim vs. maxim of Manner); the coordination model only considers hearer's economy (maxim of Quantity); the naive model considers speaker's economy only (maxim of Manner). All of these models rely on the Rational Speech Acts framework (Frank & Goodman, 2012). The model follows the implementation in Hawkins et al. (2022) designed for modelling dyadic interactions, with a modification that makes it suitable for resolving ambiguity in context.

Pragmatic model

The model consists of the set of utterances $\mathcal{U}$ and meanings $\mathcal{M}$. The set of all possible lexicons $\mathcal{L}$ is defined as all possible binary matrices of the size $|\mathcal{U}|x|\mathcal{M}|$; namely the number of utterances times the number of meanings. The lexicons that contain null messages (empty columns), are excluded from the set $\mathcal{L}$, since we do not expect the agents to say nothing when describing a meaning (this is not an option for participants in our experiment either), resulting in 32 possible lexicons. The model can be defined as follows.

The **literal listener** (1) returns a probability of inferring meaning m given lexicon l_i, utterance u and context c:

$$P_{L_0}(m|l_i,u,c) \propto \delta_{l_i|m\in[[u]]} \cdot P(m) \cdot C(m|c) \tag{1}$$

The term $C(m|c)$ corresponds to a binary function, which returns 1 if the meaning m is attested given the context c and 0 if it is not the case. $\delta_{l_i|m\in[[u]]}$ is a Kronecker delta function that evaluates to 1 when the lexicon l_i correctly associates the meaning m with the utterance u and 0 otherwise.

The **pragmatic speaker** returns a probability of producing utterance u given lexicon l_i, meaning m and context c:

$$P_{S_p}(u|l_i,m,c) \propto \exp(\alpha \cdot U(u,l_i,m,c)) \tag{2}$$

The pragmatic speaker level takes into account the literal listener's possible interpretations when calculating the expression with maximum utility given in (3), and the context that the meaning is in:

$$U(u,l_i,m,c) = \log P_{L_0}(m|l_i,u,c) - \text{cost}(u) \tag{3}$$

Moreover, as shown in (3), the pragmatic speakers also takes costs of utterances into account, by subtracting the cost value from the log probabilities obtained from the literal listener. Therefore, the trade-off between the maxim of Manner (cost) and the maxim of Quantity (maximizing utility) is implemented at the level of speaker (and listener).

The **pragmatic listener** (4) returns a probability of inferring meaning m given lexicon l_i, utterance u and context c :

$$P_{L_p}(m|l_i,u,c) \propto P_{S_p}(u|l_i,m,c) \cdot P(m) \tag{4}$$

When the agents take on the roles of either *senders* or *receivers*, they are using (2) or (4) for the encoding or decoding of messages, with the lexicon l_i sampled proportionally to $P(\mathcal{L})$. Initially, $P(\mathcal{L})$ is a uniform distribution, however after the first exchange, *senders* undergo the following update procedure after each round, implemented following Nedelcu and Smith (2022):

$$P(\mathcal{L}|m,u,c) \propto \begin{cases} P_{S_p}(u|\mathcal{L},m,c) \cdot P(\mathcal{L}) + \text{n}, \text{if TRUE} \\ \sum_{u_i \in U, u_i \neq u} P_{S_p}(u_i|\mathcal{L},m,c) \cdot P(\mathcal{L}) + \text{n}, \text{if FALSE} \end{cases} \tag{5}$$

If communication was successful (i.e chosen utterance u leads the receiver to choose the intended meaning m, *TRUE* in the equation), the prior probability of each lexicon l_i in $\mathcal{L}$ increases proportionally to how likely a pragmatic speaker is to utter u in that lexicon. This increase is adjusted with a noise term (n) to avoid sampling lexicons with zero probability, effectively preventing agents from getting stuck in a local minimum while inferring the most likely lexicon from their interactions. When the choice made by the receiver is incorrect, the probabilities of all alternative lexicon are increased instead:

$$P(\mathcal{L}|m,u,c) \propto \begin{cases} P_{L_p}(m|\mathcal{L},u,c) \cdot P(\mathcal{L}) + \text{n}, \text{if TRUE} \\ \sum_{m_i \in \mathcal{M}, m_i \neq m} P_{L_p}(m_i|\mathcal{L},u,c) \cdot P(\mathcal{L}) + \text{n}, \text{if FALSE} \end{cases} \tag{6}$$

Coordination model

This model represents a situation in which the agents do not obey the Manner maxim, i.e. when the literal speakers do not take cost into account. Overall, this model is equivalent to the pragmatic model, but the cost variable is set to zero.

Naive model

The naive model serves as a baseline for non-pragmatic communication, whereby speakers do not take context into account, and only focus in reducing cost. The **literal listener** is defined as in (1), and the **literal speaker** is defined as follows:

$$P_{S_0}(u|l_i,m,c) \propto \exp(\alpha \cdot \log(\delta_{l_i|u\in[[m]]} \cdot P(m) - cost(u))) \quad (7)$$

The literal speaker only takes into account the utterance-meaning correspondence and the cost of each utterance while choosing u. However, they do not penalize ambiguity since they do not consider the context at the listener level. The update procedures for senders and receivers remain the same with the only exception being the substitution of the pragmatic speakers and listeners with literal speakers and listeners in equations (5) and (6). Overall, this model corresponds to a scenario where there is no trade-off between the maxim of Manner and the maxim of Quantity.

Results

To determine whether the behavior of the participants can be better accounted for by the naive, pragmatic or coordination models, we generated 1000 simulations of 30 dyads playing 42 rounds of this game using each of the three models (3000 simulations in total). Each simulation was represented using the two variables that were collected for the human participants, namely the proportion of trials where long word was used with color-uninformative shape, and the proportion of trials where the short word was used with the color-informative shapes. The density plots of the simulations are represented on Figure 3 (panel A). This figure makes the differences between different pressures acting in the case of each model readily apparent. First, in the case of the naive model, most of the density is in the top left corner, which corresponds to the "only short" lexicons, meaning that only the speaker's economy is present, i.e. the aim at reducing the overall length of the message. Furthermore, the pragmatic model displays both the speaker's and the hearer's economies, whereby some lexicons are in the "only short" square, but most of them are in the top right corner, i.e. the "efficient" square. This means that both the average length of the message and the accuracy get optimized simultaneously, yielding efficient lexicons. Finally, in the case of the coordination model, the dyads are mostly equally concentrated in the "efficient" and "non-efficient" squares, meaning that only the hearer's economy is present.

The experimental results displaying the choices of lexicon by the participants are shown on Figure 3 (panel B). Most of the dyads are either in the "non-efficient" square (long word for color-informative shapes and the short word uniquely identifies the color-uninformative shape; bottom left corner) or in the "efficient" square (short word for color-informative shapes; top right corner). However, contrary to our expectations, participants are very closely balanced over these two lexicons. The human data therefore appears to most closely correspond to the behaviour of the coordination model. However, to perform this comparison more rigorously, we divided the set of simulations into the training (2400 simulations) and validation (600 simulation) sets. We then fine-tuned a Random Forest Classified to distinguish between these three types of models given the array of the two dependent variables on the training set using a 5-fold cross-validation procedure. The algorithm with the best combination of parameters has an accuracy score of 0.975 on the training set, and an accuracy score of 0.978 on the validation set. After observing the experimental data, this algorithm yields a 73% probability that the human data was produced by the coordination model (17% probability on the pragmatic model, 10% on the naive model).

Furthermore, we fitted a Bayesian logistic regression model to this data, predicting the choice of a short word given the type of stimulus (color-informative or not), with a random intercept for each dyad. The β-coefficient for the color-informative stimuli is equal to -0.34 (95 % CI: [-0.57, -0.09]), indicating a slight preference for the use long word with color-informative stimuli. We also extracted the posterior probabilities in the form P(short—color-informative) and P(long—color-uninformative) from the model, which are equal to 0.47 (95 % CI: [0.43, 0.51]) and 0.55 (95 % CI: [0.5, 0.61]), respectively. Overall, it indicates that the preference to use an efficient lexicon (short word colexifying colour-informative stimuli) is (if anything) rather small. Overall performance was above the chance level of 0.5: mean accuracy = 0.73; SD = 0.12. This suggests that participants created a successful convention, instead of simply relying on color cues and sending random messages (Figure 3, panel C).

Why would participants not show a stronger preference for efficient lexicons? One possibility we considered was that participants were overly constrained by the early conventions they formed in communication, before they fully appreciated the potential use of context to disambiguate or the efficiency cost of using long words too often (Silvey et al., 2019). To test this possibility[4], we examined whether the participants switched lexicons by comparing lexical use (long word for color-uninformative vs. short word for color-informative) in the first versus second half of the trials and in the first half of the trials (21 × 2 trials) for each dyad. We found no significant difference between the preference for word use in the first and second half of the experiment (long & color-uninformative: t(30)=1.11, p=0.28; short & color-informative: t(30)=-0.014, p=0.99). This suggests that participants tend to preserve the existing lexicons instead of switching, once one of the two lexicons ("efficient" or "non-efficient") allowing precise communication is adopted.

Discussion

We tested whether the Zipfian notions of speaker's and hearer's economies in the case of colexification could be explained by principles of pragmatic reasoning. We ran a

[4]This was a post-hoc, exploratory analysis.

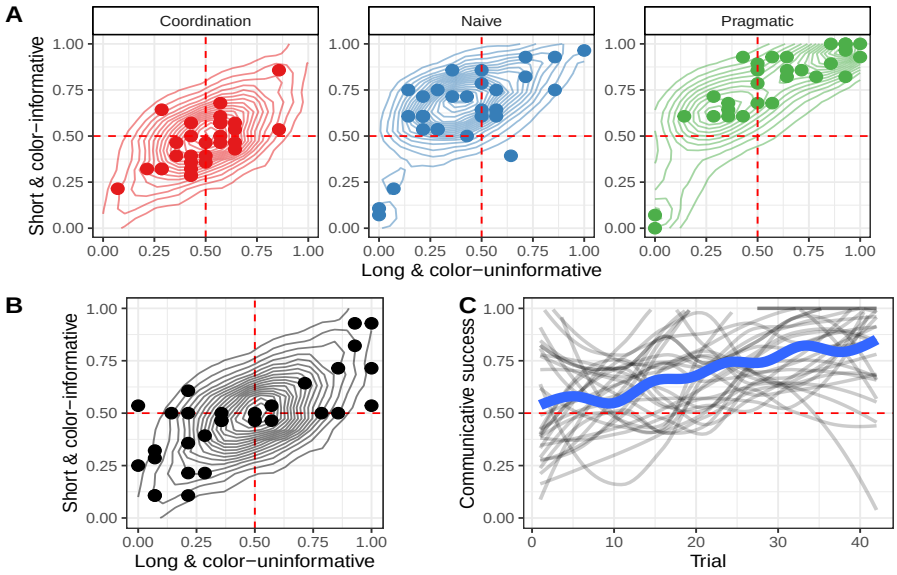

Figure 3: Summary of experimental results and model comparison. **A:** Different lexicons that the agents can converge on under different models. Kernel density estimates correspond to the distribution of 100 simulations with 30 dyads playing for 42 rounds. Scatter plots correspond to one run of the model. **B:** Lexicons used by the human participants (individual points corresponds to single dyads). The density plot corresponds to the coordination model. **C:** Convergence dynamics over trials for every dyad (grey lines). Bold blue line indicates average accuracy per trial. Dotted red line corresponds to the color-informed baseline of 50%.

preregistered experiment involving a dyadic communication game where participants were asked to communicate about geometric shapes using either short or long words. The experimental design allowed accurate communication only if the participants managed to converge on two of four possible lexicons: efficient (short word used for colexification) and non-efficient (long word used for colexification). Furthermore, we have built a set RSA-based models, where the agents were performing exactly the same task as human participants. These models allowed us to manipulate the presence or absence of speaker's and hearer's economy, resulting in three different models: a pragmatic model (speaker's economy trading off with hearer's economy), a coordination model (hearer's economy only) and a naive model (speaker's economy only). We predicted that (a) most of the participants would converge on the "efficient" lexicon, meaning that they would prefer short words when colexifying multiple meanings, and (b) the behaviour of the participants could be better explained using the pragmatic model.

Our results differ from our preregistered prediction on both counts. First, although very few dyads converged on non-productive "only short" or "only long" lexicons (showing at least that they understood the communicative nature of the task), participants in our experiment did not prefer efficient lexicons. This was further confirmed with the experimental data being most similar to the data produced by the coordination model. One possible explanation, supported by our additional exploratory analysis, was that participants were "locked-in" on a particular lexicon once the convention started to be established. This suggests that colexification can potentially occur under only relatively weak influence from semantic similarity or meaning frequency, contrary to what was hypothesized in previous literature (Xu et al., 2020; Brochhagen & Boleda, 2022). If word meanings are highly unspecified (Blutner, 1998), and the exact meaning could be enriched in particular contexts, then arbitrary colexifications are very likely to arise. Then, the preference for efficiency observed in data (Piantadosi et al., 2012) could be potentially due to weak biases acting during transmission or learning, and not communication alone (Silvey et al., 2019).

However, a more prosaic possibility is that the lack of preference for efficient lexicons lies in the design of the experiment itself. The difference in sending time that we imposed on participants might not be perceived as significantly costly. For instance, in (Kanwal et al., 2017) senders were required to press for more than 8 seconds to send the long word. Moreover, the pressure for accurate communication may be more salient in our experiment, given that participants were guaranteed a monetary bonus for each correct answer. The reward for brevity, on the other hand, was less certain, since they were promised to be rewarded only if they would be the fastest pair. Therefore, when the dyads agreed on a lexicon, they may have been reluctant to switch to another, more efficient, option, in order to maximize their profit. Overall, both of the perspectives discussed above call for a follow up experiment, where we would aim at making the speaker's and hearers economies equally salient.

References

Blutner, R. (1998). Lexical pragmatics. *Journal of semantics*, *15*(2), 115–162.

Brochhagen, T., & Boleda, G. (2022). When do languages use the same word for different meanings? the goldilocks principle in colexification. *Cognition*, *226*, 105179.

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*, 519–540.

François, A. (2008). Semantic maps and the typology of colexification. *From polysemy to semantic change: Towards a typology of lexical semantic associations*, *106*, 163.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Galantucci, B., Garrod, S., & Roberts, G. (2012). Experimental semiotics. *Language and Linguistics Compass*, *6*(8), 477–493.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Haspelmath, M. (2023). Coexpression and synexpression patterns across languages: comparative concepts and possible explanations. *Frontiers in Psychology*, *14*.

Hawkins, R. D., Franke, M., Frank, M. C., Goldberg, A. E., Smith, K., Griffiths, T. L., & Goodman, N. D. (2022). From partners to populations: A hierarchical bayesian account of coordination and convention. *Psychological Review*.

Horn, L. (1984). Towards a new taxonomy for pragmatic inference: Q-and r-based implicature. *Meaning, form and use in context*.

Horn, L. (1993). Economy and redundancy in a dualistic model of natural language. *Sky*, *1993*, 33–72.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52.

Karjus, A., Blythe, R. A., Kirby, S., Wang, T., & Smith, K. (2021). Conceptual similarity and communicative need shape colexification: An experimental study. *Cognitive Science*, *45*(9), e13035.

Lewis, M. L., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, *153*, 182–195.

Morin, O., Müller, T. F., Morisseau, T., & Winters, J. (2022). Cultural evolution of precise and agreed-upon semantic conventions in a multiplayer gaming app. *Cognitive Science*, *46*(2), e13113.

Nedelcu, V. C., & Smith, K. (2022). The complexity of a language is shaped by the communicative needs of its users and by the hierarchical nature of their social inferences. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Peloquin, B. N., Goodman, N. D., & Frank, M. C. (2020). The interactions of rational, pragmatic agents lead to efficient language structure and use. *Topics in Cognitive Science*, *12*(1), 433–445.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012, March). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291. doi: 10.1016/j.cognition.2011.10.004

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The arc nonword database. *The Quarterly Journal of Experimental Psychology: Section A*, *55*(4), 1339–1362.

Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in cognitive sciences*, *14*(9), 411–417.

Silvey, C., Kirby, S., & Smith, K. (2019). Communication increases category structure and alignment only when combined with cultural transmission. *Journal of Memory and Language*, *109*, 104051.

Xu, Y., Duong, K., Malt, B. C., Jiang, S., & Srinivasan, M. (2020). Conceptual relations predict colexification across languages. *Cognition*, *201*, 104280.

Zipf, G. K. (1945, October). The Meaning-Frequency Relationship of Words. *The Journal of General Psychology*, *33*(2), 251–256. doi: 10.1080/00221309.1945.10544509

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books.