Efficiency and Effectiveness in Task-Oriented Dialogue: On Construction Repetition, Information Rate, and Task Success

Jun Sen Yee[◇], **Mario Giulianelli**[◁], **Arabella Sinclair**[◇]
[◇]University of Aberdeen, [◁]ETH Zürich
junsen.yee@abdn.ac.uk, mgiulianelli@inf.ethz.ch, arabella.sinclair@abdn.ac.uk

Abstract

We investigate the roles that efficiency and effectiveness play in speakers' repetition of shared word sequences, or *constructions*, in task-oriented dialogue. We find that repeating constructions has negative effects on information rate and positive effects on rate of delivery, that information rate managing strategies are predictive of task success, and that this varies by the communicative function of the constructions being repeated. More effective dialogue is characterised by greater levels of shared construction usage and more efficient task-related repetition; while task-agnostic repetition can seem redundant, it can serve important efficiency and effectiveness functions. Our results provide a nuanced picture of the importance of repetition and of developing a shared lexicon for both efficiency and effectiveness in task-oriented dialogue.

Keywords: Dialogue, Information Rate, Alignment, Construction Repetition, Task-Oriented Dialogue, Task-Success, Surprisal, Efficient Communication, Speech Rate

1. Introduction

Effective interaction in task oriented dialogue requires complex, efficient strategies of coordination. Coordination can occur through alignment of speaker behaviour, from lexical and structural choice to gaze, gesture, and body posture (Brennan and Clark, 1996; Reitter et al., 2006; Holler and Wilkin, 2011; Rasenberg et al., 2020). Coordination of behaviour, whether strategic or automatic, can indicate conceptual alignment of speakers' mental models (Garrod and Pickering, 2007; Pickering and Garrod, 2004a) and interpersonal synergy between speakers (Fusaroli et al., 2014). A key aspect of coordination is speakers' tendency to repeat one another's linguistic forms. For example, ***F:*** *is it not a couple of inches* ***G:*** *aye it's a couple of inches across...* The coordination and repetition of shared linguistic forms allows speakers to collaboratively establish and maintain common ground (Pickering and Garrod, 2004b), to develop particular partner-specific language, and of that, reuse effective formulations (Brennan and Clark, 1996), and to develop a mutual understanding of their shared goals, ultimately leading to more effective communication (Wilkes-Gibbs and Clark, 1992; Ward and Litman, 2007; Friedberg et al., 2012; Reitter and Moore, 2014; Gallotti et al., 2017; Sinclair et al., 2021; Sinclair and Schneider, 2021; Norman et al., 2022).

Repetition is also strongly related to communicative efficiency. For speakers, it limits the effort required to retrieve or construct alternative realisations (e.g., Coupland, 1998; Sprenger et al., 2006), and for comprehenders, repeated sequences are more rapid to process (e.g., Bigand et al., 2005). Overall, from an information processing point of view, by exploiting the familiarity of their shared lexicon, speakers efficiently manage the information density of the language they produce (Giulianelli et al., 2022), and thus reduce their joint production and comprehension effort (Siyanova-Chanturia et al., 2017). There is a limit to how much information speakers can transmit at once before both production and comprehension costs become too high for the communication to remain robust and error-free (Shannon, 1948). Exploiting repetitions to spread information evenly, and avoid peaks in information rate can therefore be a rational strategy for marrying effectiveness and efficiency, and achieve successful communication (Genzel and Charniak, 2002; Xu and Reitter, 2017).

In the present study, we focus our attention on the relationship between efficient communication strategies and effective communication, using a navigational task-based dialogue activity (Anderson et al., 1991) as a case study. We take the repetition of shared *constructions*—multi-word configurations of structures and lexemes in the sense of Construction grammar (Tomasello, 2003; Goldberg, 2006)—as our main unit of analysis. We investigate to what extent the repetition of constructions shared between speakers contributes to *efficiency* effects, in terms of information density and delivery rate, and further explore whether these effects contribute to task *effectiveness*.

We find that repeating constructions in task-oriented dialogue contributes to information processing efficiency. We further find that construc-

LREC-COLING 2024, pages 5562–5577

tion repetition contributes to increased rate of delivery, supporting the case for a trade-off between redundancy—in terms of repetition, delivery rate, and surprisal—and duration. We find that successful dialogues are characterised by a greater level of shared construction usage, highlighting the important role of shared routines to task effectiveness (Pickering and Garrod, 2005; Norman et al., 2022), and that information mitigating strategies are important to effectiveness. Overall, our findings provide a new, nuanced picture of the importance of developing a shared lexicon to efficiency and effectiveness in task-oriented dialogue.

2. Background & Related Work

2.1. Repetition and Coordination in Dialogue

Repetitions are a natural facet of interactive language. They can be a key element in first and second language acquisition (Sinclair et al., 2021), used to minimise the consequences of false-accept errors in dialogue conversations (Giangola, 2022) as well as for clarification requests (Schlangen, 2004), and can serve to reduce the information density of the language begin produced, to minimise producer and comprehender effort (Giulianelli et al., 2022). Repetition can contribute to, or indicate speaker coordination or alignment, whereby speakers adapt and adjust their language to achieve better mutual understanding (Wilkes-Gibbs and Clark, 1992; Gallotti et al., 2017), establishing routines which can increase communicative effectiveness (Pickering and Garrod, 2005). In dialogue, repetition has been demonstrated to correlate with student learning (Ward and Litman, 2007; Sinclair et al., 2018, 2021), peer-learner collaboration quality (Sinclair and Schneider, 2021), and task success (Friedberg et al., 2012; Reitter and Moore, 2014; Norman et al., 2022). To our knowledge, the relationship between repetition, information density, and success in task-oriented dialogue remains an open question.

2.2. Processing Effort and Information Rate in Dialogue

Processing effort is a phycholinguistic construct developed as a measure of the expenditure of information-processing resources required for the perception and cognition of linguistic signals. It is typically measured in terms of neural responses or other forms of behaviour such as reading times and fixation duration. According to expectation-based psycholinguistic theories of language processing (Hale, 2001; Levy, 2008), processing cost is strongly related to the predictability of upcoming linguistic signals given their context of occurrence. An empirically successful operationalisation of predictability relies on the information-theoretic notion of *information content* or *surprisal* (Shannon, 1948). The surprisal of a signal is its negative log probability, and thus a measure of its unexpectedness. In recent years, surprisal estimates from autoregressive neural language models have been shown to be effective when predicting human comprehension behaviour in terms of reading times, gaze fixation, and neural responses (Monsalve et al., 2012; Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2020; Oh and Schuler, 2023; Michaelov et al., 2023). We thus make use of surprisal as a measure of processing effort.

Surprisal in dialogue has been found to converge between speakers (Xu and Reitter, 2018), and to depend on contextual unit and speaker role (Giulianelli et al., 2021), and speakers have been found to coordinate how information dense their language is (Xu and Reitter, 2017). Information rate can be managed making use of multiple production strategies: speakers have been found to choose words which are *less* informative when their context can predict them (Brothers and Kuperberg, 2021), and speakers' rate of delivery of highly informative content is lower (Pimentel et al., 2021).

2.3. Task Effectiveness in Dialogue

Effective task-oriented dialogues must balance efficiency with informativeness. Aylett and Turk (2004) argue that for robust, clear, and easy to interpret communication, there is an inverse relationship between redundancy and duration, hypothesising that robustness is improved by spreading information more evenly. Asynchrony between speakers' fluctuations in information rate—their taking turns to contribute highly information dense content—has been found predictive of task success (Xu and Reitter, 2017). Alignment—either automatic (Pickering and Garrod, 2004a) or strategic (Garrod and Pickering, 2007)—measured via between-speaker overlap and repetition, has been found to occur to a higher degree in more successful dialogues (Wilkes-Gibbs and Clark, 1992; Gallotti et al., 2017; Reitter and Moore, 2014; Sinclair and Schneider, 2021), specifically when this repetition pertains to the task (Ward and Litman, 2007; Friedberg et al., 2012; Norman et al., 2022).

3. Data

3.1. Corpus Choice

As a corpus of task-oriented dialogue, we select the HCRC Map Task Corpus[1] (Anderson et al., 1991).

[1] https://groups.inf.ed.ac.uk/maptask/

	Mean±Std	Median	Min	Max
Dialogue				
Length (u)	211±107	183	42	686
Length (w)	1201±648	1014	266	4515
Duration (s)	408±186	363	156	1140
Utterance				
Length (w)	5±6	3	1	106
Duration (s)	1.52±1.96	0.82	0.05	44.68
Task Success				
PATHDEV	71±49	56	4	227

Table 1: Map Task statistics reported in utterances (u), words (w), and seconds (s). PATHDEV is the path deviation score: the distance between gold route and attempt.

We study this corpus in detail, with the aim of discovering whether speaker use of construction repetition is an *efficient* and *effective* communication strategy. The Map Task Corpus is comprised of 128 task-oriented dialogues, each conducted by two participants: an instruction *giver*, and an instruction *follower*. Participants were given similar, but non-identical, maps populated with imaginary landmarks.[2] The instruction *giver* was provided with a map that contained a start point, pre-defined route, and finish point, while the instruction *follower* was provided only the start point. The task required the *follower* to reproduce the route on their own map, as dictated by the *giver*. Conversations were recorded and transcribed.

3.2. Corpus Measures of Effectiveness and Efficiency

Two key aspects of task-oriented dialogue are whether it is *effective*, i.e., whether it leads to task success, and whether the linguistic strategies used by speakers to complete the task are *efficient*.

Effectiveness: Task Success The goal of the Map Task was for the instruction giver to provide instructions sufficient for the instruction follower to replicate a specific path traversing their shared map without having seen it. The performance of the task was measured in terms of the deviation of the route that the *follower* drew on their map from the pre-defined route on the *giver's* map (PATHDEV). A larger value of PATHDEV indicates a poorer task performance. Effectiveness, or task success, is therefore measured as the inverse of PATHDEV.

Time Efficiency: Rate of Delivery Efficiency can be measured as a function of the speed at which speakers communicate their information, or speech rate in tokens per second (Giulianelli et al., 2022; Reitter et al., 2006), or as the inverse, i.e., average token duration (Pimentel et al., 2021). We choose the latter, making use of *token duration*, or rate of delivery (in terms of seconds per token) as a measure of time efficiency. We measure average token duration across different spans of the text, at the level of the entire dialogue, individual utterances, and constructions (Sections 5 and 6).

3.3. Extracting Repeated Constructions

The main focus of this study is on repeated *constructions*, which are multi-word sequences repeated within a dialogue.

	Mean±Std	Median	Max
Constr. Length (w)	3.43±0.79	3	9
Constr. Frequency	4.84±2.71	4	19
Constr. Rep. Dist. (w)	892±1179	424	8940
Constr. Incidence	102±72	80	282
— *Landmark*	22±19	17	70
— *Direction*	45±33	35	143
— *Generic*	37±29	27	117
Utterance Length (w)	5.66±6.16	4	68

Table 2: Construction (Constr.) statistics. The reported statistics are for the analysis split, described in Section 4.2. Length and repetition distance (Rep. Dist.) measured in words (w).

To extract repeated constructions from the Map Task Corpus, we use *dialign*, a framework for sequential pattern mining (Dubuisson Duplessis et al., 2017).[3] We then apply several filtering conditions to the extracted lexicons of constructions. Specifically, we discard repeated constructions with fewer than three tokens, or those repeated fewer than three times. Repeated constructions consisting solely of punctuation or of more than half filled pauses are also excluded.[4] The remaining constructions are what we consider as *constructions*. Relevant construction statistics can be found in Table 2.

Dialogue specificity of constructions We measure dialogue specificity using Pointwise Mutual Information (PMI). PMI measures how much more likely an construction is to occur in a specific dialogue than it is to occur in general. In other words, it measures the specificity of a construction to the dialogue. PMI is computed as follows:

$$PMI(c, d) = \log_2 \frac{P(c|d)}{P(c)} \quad (1)$$

[2]The mismatch between two maps may be in terms of landmarks' existence, number of appearance, name, or location.

[3]https://github.com/GuillaumeDD/dialign#framework

[4]The full list of filled pauses can be found in Appendix B.

A higher PMI for a construction indicates that it is more strongly associated with, or specific to, the dialogue due to its increased frequency of occurrence within the dialogue compared to its general usage. Conversely, a PMI lower than 1 indicates that the construction is not specific to the dialogue, and can be seen as dialogue-independent.

Communicative functions of constructions We distinguish constructions by their communicative function, differentiating between two types of task-specific construction: those relating to the *landmarks* within the task (e.g., *the gold mine, the extinct volcano*), and those relating to *directions* being used by participants to complete the task (e.g., *to the left, straight up to*).

The remaining constructions we consider to be *generic*. These are typically non-referential constructions (since they relate to other, less topic- or task-determined functions) such as discourse organisers or stance constructions (Biber et al., 2004; Biber and Barbieri, 2007) (e.g., *now if you, what I mean, oh right okay*) The vocabulary lists constructed to perform this functional categorisation can be found in Appendix F.1.3. Table 3 contains examples.

Landmarks	***Directions***	***Generic***
the rope bridge	the left-hand side	do you have
the white mountain	about an inch	there's a
the diamond mine	of the page	I've got a
the gold mine	to the right	have you got
the trout farm	the top of	I've got
the slate mountain	side of the	have you got a
of the mountain	about three inches	do you have a
the dead tree	to the left	you're at
of the stile	the top of the	until you're
the carved stones	to your left	I don't

Table 3: Examples of the 10 most frequent constructions for each communicative function.

4. Estimating Information Processing Efficiency

We use surprisal estimates as approximate measurements of the cost necessary to process words and utterances (Levy, 2008; Smith and Levy, 2013; Goodkind and Bicknell, 2018). We calculate surprisal using an autoregressive neural language model equipped with a simple utterance-level, and dialogue-specific, adaptation mechanism (van Schijndel and Linzen, 2018; Giulianelli et al., 2022).

4.1. Measures of Surprisal

The surprisal of a word w_i is the negative logarithm of the word probability given the utterance so far $u_{:wi}$ and the local dialogue context l:

$$S(w_i|u_{:w_i}, l) = -\log_2 P(w_i|u_{:w_i}, l) \quad (2)$$

Following Giulianelli et al. (2022), we determine the size of the local dialogue context l by computing the average word-level speech rate of the corpus and multiplying it by 15 seconds.[5] The resulting context size for Map Task is 65 words.

Utterance surprisal To compute the surprisal of an entire utterance, we average its per-word surprisal values:[6]

$$S(u; l) = \frac{1}{|u|} \sum_{w_i \in u} S(w_i|u_{:w_i}, l) \quad (3)$$

Construction surprisal We apply the same averaging strategy to compute the surprisal of a construction:

$$S(c; u_{:c}, l) = \frac{1}{|c|} \sum_{w_i \in c} S(w_i|u_{:c}, l) \quad (4)$$

Facilitating effect To quantify the change in surprisal contributed by a construction to its containing utterance, we calculate its *facilitating effect*, i.e., the logarithm of the ratio between construction surprisal and the surprisal of the utterance context (Giulianelli et al., 2022):

$$FE(c; u, l) = \log_2 \frac{\frac{1}{|u|-|c|} \sum_{c \not\ni w_j \in u} S(w_j|u_{:w_i}, l)}{\frac{1}{|c|} \sum_{w_i \in c} S(w_i|u_{:c}, l)} \quad (5)$$

Facilitating effect is positive when the surprisal of a construction is lower than the surprisal of its utterance context, it is set to zero when they are equal, and it is negative otherwise.

4.2. Language Model

To estimate surprisal (Eq. 2), we use GPT-2 (Radford et al., 2019), a Transformer-based autoregressive language model. Its estimates have been shown to be predictive of comprehension behaviour (e.g., Wilcox et al., 2020; Shain et al., 2024). To obtain estimates for the full Map Task corpus, we iteratively fine-tune GPT-2 on an 80% split of the data, then estimate surprisal over the remaining 20% split, and repeat this for 5 distinct splits. We

[5] 15 seconds are the the locus of local repetition effects identified in previous work (Reitter et al., 2006).

[6] The best aggregation depends on the type of comprehension behaviour surprisal is intended to predict (see, e.g., Meister et al., 2021). We use averaging because it captures a notion of information *rate*, rather than cumulative cost, and is less affected by utterance length (Keller, 2004).

thus collect surprisal estimates for the full corpus while ensuring that the model has not been fine-tuned on the target dialogues. [7] Transcribed spoken dialogue only makes up a very small part of the GPT-2 pre-training corpus, so fine-tuning is a first necessary domain adaptation step (as shown by the large decrease in perplexity after fine-tuning; see Appendix C).

We equip the language model with a simple adaptation mechanism that allows the model to continually learn from exposure to utterances within a given dialogue. After each utterance is processed, we perform a back-propagation step using a cross-entropy loss to update the parameters of the language model. This mechanism makes the model more cognitively plausible, and yields surprisal estimates that are more in line with human expectations (van Schijndel and Linzen, 2018). To determine an appropriate learning rate for the adaptation mechanism, we follow Giulianelli et al. (2022) and select the learning rate which yields an optimal combination of in-distribution and out-of-distribution generalisation (Hupkes et al., 2022) in a battery of cross-validation tests. Further details in Appendix D.

5. Efficiency of Repetition

We explore efficiency through examining the relationship between construction repetition, information density, and rate of delivery. For our statistical analyses, we employ t-tests and linear mixed-effect models. In the mixed effect models, we always include construction length and repetition index within the current turn as baseline predictors (baseline model), as well as dialogue and speaker as random effects to capture group-level variability.[8]

5.1. Construction Repetition Facilitates Processing in Task-Oriented Dialogue

We analyse properties of construction surprisal and facilitating effect (FE) in task-oriented dialogue, studying the effects of construction repetition on utterance levels of information rate.

Construction use reduces information rate. We compare the surprisal profiles of construction and non-construction sequences. Our first hypothesis is that the surprisal of constructions will be lower. This would confirm, also from an information-theoretic angle, that constructions have a processing advantage (Conklin and Schmitt, 2012; Carrol and Conklin, 2020). Our second hypothesis is that the facilitating effect of constructions will be positive. This would indicate that they have a mitigating effect on utterance surprisal (Giulianelli et al., 2022). Both hypotheses are confirmed: constructions have 39% lower surprisal ($t = -35.943$, $p < 0.05$, $95\%\ CI = [-1.724\!:\!-1.596]$), and approximately 6 times higher FE than non-construction word sequences ($t = 17.917$, $p < 0.05$, $95\%\ CI = [0.672\!:\!0.784]$).[9] Taken together, these results constitute new empirical evidence that construction use reduces information rate in task-oriented dialogue.

Construction repetition has facilitating effects. We then analyse construction *repetitions*, hypothesising that they have higher facilitating effect than first construction mentions. We find the FE of construction repetitions is 57% higher than that of first mentions ($t = -19.796$, $p < 0.05$, $95\%\ CI = [-0.415 : -0.367]$). These strong utterance surprisal reduction effects suggest that repetition substantially contributes to making interactions more cost-efficient.

Facilitating effects of construction repetition are cumulative and decay. Given prior findings for *open-domain* dialogue, we anticipate two further key traits of the FE associated with constructions: that the facilitating effects will decay as the distance between subsequent mentions increases and accumulate over subsequent repetitions (Giulianelli et al., 2022). We confirm empirically that these trends hold also in task-oriented dialogues. FE is lower as the distance between the current repetition and its previous mention increases ($\beta = -0.0.149$, $p < 0.05$, $95\%\ CI = [-0.160\!:\!-0.138]$), and higher as the number of mentions of a construction increases ($\beta = 0.161$, $p < 0.05$, $95\%\ CI = [0.130 : 0.192]$). Including these two predictors results in improved model fit over the baseline model. These results indicate a clear relationship between proximity and cumulativity of repetitions, and enhanced cost-efficiency.

5.2. A Construction's Communicative Function Affects Repetition Efficiency

We now group constructions according to their communicative function, distinguishing between *landmarks*, *directions*, and *generic* constructions. rere 1 shows construction FE by communicative function (see Section 3.3 and Table 3 for example constructions).

Task-related repetition shows higher facilitating effect. Our main expectation is that *landmarks*

[7] More details on fine-tuning in Appendix C.

[8] Full model output for the results in this section can be found in Appendices F.1 (Section 5.1), F.1.3 (Section 5.2), and F.2 (Section 5.3).

[9] The method to extract non-construction sequences can be found in Appendix E.

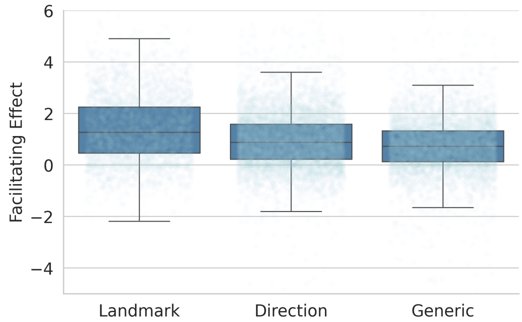

Figure 1: FE by communicative function.

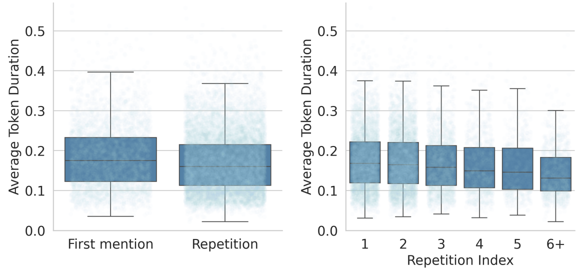

Figure 2: Duration of constructions decreases with subsequent repetitions

and *directions* will exhibit higher facilitating effect in line with trends observed for referential constructions in conversational settings (Giulianelli et al., 2022). Our analysis confirms this to be the case: the FE of repeated *landmarks* is significantly higher than that of the other two construction types, and the FE of *directions* is higher than that of *generic* constructions ($landmarks$: $\beta = 0.634$, $p < 0.05$, $95\%\ CI = [0.576 : 0.691]$; $directions$: $\beta = 0.195$, $p < 0.05$, $95\%\ CI = [0.148 : 0.242]$).[10] We also observe improved model fit when including these factors. Our finding of stronger FE in landmarks and directions gives one possible, information-theoretic, explanation for the importance of repeating task-related routines to collaboration as shown in prior work (Ward and Litman, 2007; Friedberg et al., 2012; Norman et al., 2022).

5.3. Repetition and Rate of Delivery

Another dimension of efficiency that interacts with information rate and repetition is the rate of delivery. We expect repeated constructions to have a higher speed of delivery, and that more surprising utterances and constructions will be delivered more slowly (Aylett and Turk, 2004; Pimentel et al., 2021). We define our normalised measure of $duration$ as $\frac{duration_s}{token_s}$, where s (for 'sequence') is either an utterance or a construction. To identify factors influencing utterance and construction duration, we fit linear mixed effect models with duration as a response variable, and always include construction length, average per-word character length and average word frequency as baseline predictors, as well as dialogue and speaker as random effects.[11]

The more redundant the repetition, the higher the speed of delivery. We find that the duration of constructions is shorter than that of non-constructions ($t = -24.869$, $p < 0.05$, $95\%\ CI = [-0.048 : -0.043]$) and that repetitions take less time than first mentions ($t = -9.517$, $p < 0.05$, $95\%\ CI = [-0.016 : -0.012]$). For repetitions, we observe that duration further decreases with the number of repetitions ($\beta = -0.061$, $p < 0.05$, $95\%\ CI = [-0.073 : -0.049]$) and slightly decays with a repetition's distance from the previous mention ($\beta = 0.012$, $p < 0.05$, $95\%\ CI = [0.008 : 0.016]$). We observe clear differences between construction communicative function: *landmarks* and *directions* take longer to produce than generic constructions ($Landmarks$: $\beta = 0.360$, $p < 0.05$, $95\%\ CI = [0.328 : 0.391]$, $Directions$: $\beta = 0.410$, $p < 0.05$, $95\%\ CI = [0.391 : 0.429]$. Overall, these results provide further evidence that redundant information is delivered more rapidly (Aylett and Turk, 2004). However, the slower rate of task-related repetitions suggests that expressions whose function is essential to task completion may require additional emphasis for effective dialogue.

Surprisal duration trade-off coexists with extended Facilitating Effect. Further evidence for the relationship between redundancy and duration comes from surprisal theory (Hale, 2001; Levy, 2008). We expect to find a surprisal-duration trade-off, as found by Pimentel et al. (2021): whereby in delivering unexpected linguistic units more slowly, speakers spread information more evenly over time. We indeed find that higher construction surprisal corresponds to longer duration, and thus lower speed of delivery ($\beta = 0.045$, $p < 0.05$, $95\%\ CI = [0.037 : 0.054]$). We also find that higher FE corresponds to longer duration ($\beta = 0.021$, $p < 0.05$, $95\%\ CI = [0.011 : 0.031]$). Since FE correlates inversely with surprisal ($r = -0.484$) we might have expected a negative effect of FE on duration. Instead, the observed moderate positive effect can be interpreted as a strategy to stretch FE over longer time spans, and it may coincide with emphasis points. The fact that FE and surprisal are independently predictive of duration further suggests that speakers may use two information rate mitigation

[10]These linear mixed effect models include *landmarks* and *directions* as categorical fixed effects, with *generic* included in the intercept.

[11]We obtain word frequencies extracted from a corpus of subtitles by Brysbaert et al. (2012). More details and full results in Appendix F.2.

strategies: distributing information evenly over their production, and introducing short intervals of further reduced information load (via extending the facilitating effects with lower delivery rate).

6. Analysing Effectiveness: Efficient Strategies for Task Success

We now explore the *efficiency* effects found in Section 5 with respect to their *effectiveness* in task-oriented dialogue. We define *task success* as the normalised inverse of the path deviation score PATHDEV (see Section 3.2). This will be the dependent variable in our analyses. As predictors of task success, we consider factors capturing different aspects of efficient language use: surface-level patterns of repetition, information rate, and duration.

Construction usage and repetition: we include construction usage proportion (measured as the share of dialogue tokens belonging to constructions), the number of unique constructions, the average construction length, the construction frequency, and the distance between repeated constructions as predictors. These factors capture efficiency in terms of raw amounts of construction usage as well as in terms of patterns of repetition (e.g., singling out local repetitions). We expect task success to be predicted by such surface-level patterns of efficient language use.

Information rate: our information-theoretic predictors are facilitating effect, surprisal (average over constructions and over utterances), variance in surprisal (over the whole dialogue as well as within utterances), as well as the slope of surprisal and facilitating effect over the course of a dialogue. On the one hand, higher facilitating effect and lower surprisal correspond to lower processing effort, and can thus be expected to predict success. The same reasoning holds for surprisal and facilitating effect slopes, which capture trajectories of information rate over the course of dialogues. At the same time, excessively low information rates may be a signature of uninformative interactions. We thus expect to observe strategies that trade off efficiency and informativeness to be predictive of success. Our predictors also include surprisal variance, which is an estimate of global (when measured over the dialogue) and local (within the utterance) uniformity of information. Under the uniform information density hypothesis (UID), which postulates information transmission is optimal when abrupt changes in information are avoided (Genzel and Charniak, 2002; Levy and Jaeger, 2007), we would expect dialogues with higher uniformity to be more successful.

Duration: we consider dialogue length as measured by total dialogue tokens as well as total dialogue duration in seconds. We expect longer dialogues may be more successful as this can indicate more information transmitted and a greater level of detail in instruction giving and following utterances.

With the factors presented above included as fixed effects, we fit a linear model that predicts task success. For each model, we perform iterative step-wise removal of factors with the lowest contribution until only factors significantly predictive of task success remain.[12] In our first model, which does not distinguish between self-repetitions and repetitions of shared constructions, nor between constructions with different communicative functions or uttered by dialogue participants with different speaker roles, only the total number of tokens is identified as a significant predictor of task success. In the next sections, we consider the predictors described above split by communicative function (Section 6.1), whether constructions are shared or unshared (Section 6.2), and speaker role (Section 6.3).

6.1. Communicative Function

We split factors that relate to construction usage and repetition by a construction's communicative function: *landmark*, *generic*, or *direction*. We expect task-related repetition (i.e., *landmark* and *generic*) to be important to success (Norman et al., 2022). We find that effective dialogues are characterised by speakers using less surprising constructions referring to landmarks ($\beta = -1.442$, $p < 0.05$, $95\%\ CI = [-2.267 : -0.616]$) and repeating direction constructions with increasing facilitating effect (FE slope: $\beta = 1.522$, $p < 0.05$, $95\%\ CI = [0.021 : 3.022]$). These information rate management strategies for task-related constructions indicate that processing efficiency is an important predictor for task success. Finally, effective generic constructions are more surprising ($\beta = 1.557$, $p < 0.05$, $95\%\ CI = [0.305 : 2.809]$) and have higher facilitating effect ($\beta = 1.679$, $p < 0.05$, $95\%\ CI = [0.366 : 2.992]$), indicating that while also facilitating, these task-agnostic constructions can be effective and informative.

6.2. Shared vs. Unshared

We then fit separate models for shared (i.e., repeated by both speakers) vs. unshared (repeated by a single speaker) constructions. On the one hand, shared constructions can attest to the development of shared routines, which allow speakers to establish a common ground necessary for attaining their joint goals. On the other hand, self-repetitions are related to processing efficiency in language production, as a recently used construction can be more readily accessible for

[12] Model details can be found in Appendix G.

the speaker. We thus expect different efficient strategies to be predictive of task success for shared vs. unshared constructions.

We observe for both models that within-utterance uniformity is *not* a characteristic of successful dialogue; more successful dialogues are characterised by higher surprisal variance (Unshared: $\beta = 0.732$, $p < 0.05$, $95\%\ CI = [0.032 : 1.431]$, Shared: $\beta = 1.626$, $p < 0.05$, $95\%\ CI = [0.264 : 2.988]$). This result contributes to the body of evidence (Giulianelli and Fernández, 2021; Giulianelli et al., 2022) indicating that strategies of information transmission in dialogue are not as rational, or as optimal, as UID would predict.

Unshared We find that unshared landmark constructions in more effective dialogues are less surprising ($\beta = -1.080$, $p < 0.05$, $95\%\ CI = [-1.873 : -0.288]$), longer ($\beta = 1.412$, $p < 0.05$, $95\%\ CI = [0.700 : 2.125]$), and less frequently repeated ($\beta = -1.382$, $p < 0.05$, $95\%\ CI = [-2.50 : -0.260]$). Longer, less surprising references to landmarks may be easier to resolve, and thus less likely to need to be repeated. Furthermore, using common constructions for giving directions is more effective than using highly dialogue-specific ones (PMI: $\beta = -0.708$, $p < 0.05$, $95\%\ CI = [-1.273 : -0.143]$), more so when their use becomes more facilitating (FE slope: $\beta = 2.117$, $p < 0.05$, $95\%\ CI = [0.995 : 3.239]$). We take this to indicate that effective unshared directions should not diverge too much from common vocabulary, likely already familiar to and thus clear to both speakers. Finally, more successful generic constructions become more surprising over time (S slope, $\beta = 1.440$, $p < 0.05$, $95\%\ CI = [0.27 : 2.603]$), which suggests that the information rate of task-agnostic repetitions is less important to effectiveness.

Shared We find the proportion of shared constructions within the dialogue is an important predictor of effectiveness ($\beta = 2.575$, $p < 0.05$, $95\%\ CI = [0.978 : 4.172]$). In other words, in line with our predictions, the development and reuse of *shared* routines is a characteristic of successful dialogue. Effective landmark use consists of speakers sharing a smaller set of constructions ($\beta = -1.945$, $p < 0.05$, $95\%\ CI = [-3.179 : -0.710]$) that are shorter ($\beta = -1.199$, $p < 0.05$, $95\%\ CI = [-2.366 : -0.031]$) and less surprising ($\beta = -6.182$, $p < 0.05$, $95\%\ CI = [-9.106 : -3.258]$). We also observe that lower FE of landmark constructions corresponds to more effective dialogues ($\beta = -2.359$, $p < 0.05$, $95\%\ CI = [-4.340 : -0.377]$)—which could be to do with both speakers making short confirmatory repetitions (when a construction fills the whole utterance, its FE is equal to 0; see Section 6.3 for an example)—and so does a positive surprisal slope ($\beta = 2.277$, $p < 0.05$, $95\%\ CI = [0.207 : 4.347]$), indicating an increasing trajectory of information rate over time. Furthermore, we find effective direction constructions to be longer ($\beta = 3.089$, $p < 0.05$, $95\%\ CI = [0.309 : 5.869]$), fewer (inventory size: $\beta = -2.231$, $p < 0.05$, $95\%\ CI = [-4.173 : -0.289]$), and not repeated locally (distance between repetitions: $\beta = 2.515$, $p < 0.05$, $95\%\ CI = [0.942 : 4.087]$). Unlike unshared direction constructions, more surprising directions are effective if they are repeated by both speakers ($\beta = 2.129$, $p < 0.05$, $95\%\ CI = [0.582 : 3.677]$). Finally, shared generic constructions are effective when frequently re-used ($\beta = 1.982$, $p < 0.05$, $95\%\ CI = [0.096 : 3.868]$), repeated over longer distances ($\beta = 2.286$, $p < 0.05$, $95\%\ CI = [0.467 : 4.104]$), more surprising with re-use (S slope: $\beta = 1.876$, $p < 0.05$, $95\%\ CI = [0.483 : 3.269]$), and when they have lower FE ($\beta = -1.327$, $p < 0.05$, $95\%\ CI = [-2.524 : -0.130]$). While these properties do not correspond to processing efficiency, they may serve the purpose of allowing speakers to establish rapport (Cappella, 1990) and contributing to common ground, which our measures do not directly capture.

6.3. Speaker Role

Finally, we investigate effective construction usage patterns specific to speaker role. Followers repeating fewer unique landmarks ($\beta = -1.230$, $p < 0.05$, $95\%\ CI = [-2.152 : -0.308]$) that are not highly specific or unique to that particular dialogue (PMI: $\beta = -1.622$, $p < 0.05$, $95\%\ CI = [-2.517 : -0.727]$) are characteristic of successful dialogues. We interpret this—combined with instruction givers repeating more unique landmarks ($\beta = 1.015$, $p < 0.05$, $95\%\ CI = [0.173 : 1.857]$)—as indicating the importance of the giver describing landmarks clearly, such that the follower will not repeat them. We also observe that successful landmark use by the giver shows increasing facilitating effects over time ($\beta = 2.236$, $p < 0.05$, $95\%\ CI = [0.647 : 3.826]$), indicating better integration of those constructions into the common ground (even though the giver may need to repeat themselves, this can lead to more successful grounded communication). Frequent direction repetition by followers is effective ($\beta = 2.376$, $p < 0.05$, $95\%\ CI = [0.372 : 4.381]$); the opposite is true for instruction givers ($\beta = -0.920$, $p < 0.05$, $95\%\ CI = [-1.820 : -0.020]$), for whom repeating the same direction multiple times indicates unresolved misunderstanding. We also find that followers in more successful dialogues repeat direction constructions with lower surprisal ($\beta = -1.909$, $p < 0.05$, $95\%\ CI = [-3.595 : -0.224]$), and lower FE ($\beta = -2.660$, $p < 0.05$, $95\%\ CI = [-4.337 : -0.983]$). This is a combination particularly common for utterances consisting only of a repeated construction, often indicating a clarification or confirmatory repetition such as G: *er there's an avalanche just slightly to the right* F: *to the right*. This result thus highlights the importance of clarification requests and confirmatory repetitions to task success. Finally, local repetition of generic constructions ($\beta = -1.247$, $p < 0.05$, $95\%\ CI = [-2.311 : -0.182]$) that have a facilitating effect ($\beta = 1.774$, $p < 0.05$, $95\%\ CI = [0.594 : 2.953]$) is a successful strategy for followers. For instruction givers, effective generic construction use consists of re-using longer ($\beta = 1.757$, $p < 0.05$, $95\%\ CI = [0.207 : 3.307]$), more surprising constructions ($\beta = 1.810$, $p < 0.05$, $95\%\ CI = [0.575 : 3.045]$) that increase in facilitating effect ($\beta = 7.336$, $p < 0.05$, $95\%\ CI = [2.477 : 12.194]$). This can be viewed as a strategy that trades off efficiency and effectiveness: including more information, at a higher information rate, while controlling for processing cost.

7. Discussion & Conclusion

Repeating information during linguistic interactions may seem redundant and, as such, not conducive to efficient or effective communication. Our in-depth analysis of English dialogues in the Map Task navigational game shows, instead, that repetition of lexicalised constructions contributes to efficiency at multiple levels, and that this efficient hidden side of repetition serves a key function in task-effective dialogue.

Estimates of utterance and construction surprisal, which we take as proxies for processing cost, show that construction repetition facilitates information processing in task-oriented dialogue: surprisal is lowered, and facilitating effect increased by: (i) construction use, (ii) construction repetition, (iii) repetition frequency, and (iv) distance from the previous mention. Our results contribute new evidence, complementary to previous work in open-domain dialogue (Giulianelli et al., 2022), that speakers use efficient information-transmission strategies in dialogue, even when these are not the most optimal within the noisy channel model of communication. We find speakers in Map Task distribute information evenly over time, as predicted by information-theoretic and psycholinguistic accounts of language production (Genzel and Charniak, 2002; Levy and Jaeger, 2007), but this uniform distribution of information does *not* positively impact task effectiveness. By repeating constructions with high facilitating effect, speakers produce short bursts of non-uniform, reduced information load. This points to uniformity being a base-level rational strategy of use of the communication channel, on top of which speakers show a tendency for cost reduction—a behaviour which seems to be specific to dialogic interactions (Giulianelli and Fernández, 2021; Giulianelli et al., 2021, 2022). When considering the temporal dimension of efficiency, we find repeated constructions have a *higher rate of delivery* and that higher rate of delivery corresponds to lower surprisal. These results provide new support for the existence of a trade-off between duration and redundancy—which results in more uniform information distribution—both in terms of repetitions (Aylett and Turk, 2004) and surprisal (Pimentel et al., 2021). However, the rate of delivery of constructions does not appear to directly influence effectiveness. Beyond uniformity, we do find that the information profile of constructions is an important factor in predicting successful dialogues: efficient task-specific routines and more information dense generic ones are most effective. We view this as speakers combining the efficiency and effectiveness of routines to help them achieve successful communication (Garrod and Anderson, 1987; Brennan and Clark, 1996; Pickering and Garrod, 2005).

We find that higher levels of *shared* construction usage is predictive of successful dialogues. The re-use of shared constructions allows speakers to create short routines (Pickering and Garrod, 2005), induces alignment (Garrod and Pickering, 2009), reduces information load (Siyanova-Chanturia et al., 2017), and critically, as we observe, leads to predictably more successful dialogues. We find the *communicative function* of a construction serves an important role, both in the construction's information profile and in its contribution to successful dialogue, in line with previous findings (Sprenger et al., 2006; Tremblay et al., 2011). While task-related repetition is important to success, so too is establishing shared routines that include generic language. These may serve important dialogue facilitating and social functions, with sharing language being a sign of developing rapport (Cappella, 1990; Sinha and Cassell, 2015). Finally, we find successful strategies vary by speaker role. For example, repeating directions is an effective strategy for the follower, but not the giver.

We suspect that while some of our findings may be task-dependent, information rate mitigation and shared construction usage are general communication strategies at play in dialogue. We thus look forward to seeing these findings replicated across a wider range of task-oriented dialogues. Overall, our results, in line with prior work (Pickering and Garrod, 2005; Friedberg et al., 2012; Reitter and Moore, 2014; Sinclair et al., 2021; Norman et al., 2022) contribute new information about the important role of construction repetition in the efficiency and effectiveness of task-oriented dialogue.

Acknowledgements

We thank the anonymous LREC-COLING reviewers for their thoughtful questions and suggestions, as well as the SemDial reviewers who provided valuable feedback on a previous version of this paper.

Limitations

Limitations to our work include the size of the corpus: replicating these results on a larger dataset can provide more robust evidence for the effects we report. Since we only consider a single type of dialogue task, the findings we report relate to specific types of constructions, which may not generalise across different tasks, for example, in a task where speakers have to agree on referring expressions to a less concrete landmark (e.g., in a real-world map setting) the relative importance of different construction functions may change. Finally, our results are limited to English-language dialogue, it would be very interesting to compare which properties of successful communication generalise across languages. We hope future work will try replicating our findings on different task-oriented dialogue settings, in particular with respect to the differences between task-specific vs. more generic types of repetition.

References

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

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, 47(1).

Douglas Biber and Federica Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for specific purposes*, 26(3).

Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. If you look at...: Lexical bundles in university teaching and textbooks. *Applied linguistics*, 25(3).

Emmanuel Bigand, Barbara Tillmann, Bénédicte Poulin-Charronnat, and D Manderlier. 2005. Repetition priming: Is music special? *The Quarterly Journal of Experimental Psychology Section A*, 58(8).

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

Trevor Brothers and Gina R. Kuperberg. 2021. Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, 116:104174.

Marc Brysbaert, Boris New, and Emmanuel Keuleers. 2012. Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, 44:991–997.

Joseph N Cappella. 1990. On defining conversational coordination and rapport. *Psychological Inquiry*, 1(4):303–305.

Gareth Carrol and Kathy Conklin. 2020. Is all formulaic language created equal? Unpacking the processing advantage for different types of formulaic sequences. *Language and Speech*, 63(1).

Kathy Conklin and Norbert Schmitt. 2012. The processing of formulaic language. *Annual Review of Applied Linguistics*, 32.

Nikolas Coupland. 1998. Smooth talkers: The linguistic performance of auctioneers and sportscasters.

Guillaume Dubuisson Duplessis, Chloé Clavel, and Frederic Landragin. 2017. Automatic measures to characterise verbal alignment in human-agent interaction. In *SIGdial*.

Heather Friedberg, Diane Litman, and Susannah BF Paletz. 2012. Lexical entrainment and success in student engineering groups. In *2012 IEEE Spoken Language Technology Workshop*. IEEE.

Riccardo Fusaroli, Joanna Rączaszek-Leonardi, and Kristian Tylén. 2014. Dialog as interpersonal synergy. *New Ideas in Psychology*, 32.

M. Gallotti, M.T. Fairhurst, and C.D. Frith. 2017. Alignment in social interactions. *Consciousness and Cognition*, 48.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2).

Simon Garrod and Martin J Pickering. 2007. Alignment in dialogue. *The Oxford handbook of psycholinguistics*.

Simon Garrod and Martin J Pickering. 2009. Joint action, interactive alignment, and dialog. *Topics in Cognitive Science*, 1(2).

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of ACL*.

James Giangola. 2022. You can say that again: The role of repetition in conversation design.

Mario Giulianelli and Raquel Fernández. 2021. Analysing human strategies of information transmission as a function of discourse context. In *Proceedings of CoNLL*.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2021. Is information density uniform in task-oriented dialogues? In *Proceedings of EMNLP*.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2022. Construction repetition reduces information rate in dialogue. In *Proceedings of AACL-IJCNLP*.

Adele Goldberg. 2006. Constructions at work: Constructionist approaches in context.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of NAACL*.

Judith Holler and Katie Wilkin. 2011. Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior*, 35.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. 2022. State-of-the-art generalisation research in nlp: A taxonomy and review. *arXiv preprint arXiv:2210.03050*.

Frank Keller. 2004. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proceedings of EMNLP*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3).

Roger Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of EMNLP*.

James A Michaelov, Megan D Bardolph, Cyma K Van Petten, Benjamin K Bergen, and Seana Coulson. 2023. Strong prediction: Language model surprisal explains multiple n400 effects. *Neurobiology of Language*.

Irene Fernandez Monsalve, Stefan L Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of EACL*.

Utku Norman, Tanvi Dinkar, Barbara Bruno, and Chloé Clavel. 2022. Studying alignment in a collaborative learning activity via automatic methods: The link between what we say and do. *Dialogue & Discourse*, 13(2).

Byung-Doh Oh and William Schuler. 2023. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *TACL*, 11.

Martin J. Pickering and Simon Garrod. 2004a. The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences*, 27(2).

Martin J Pickering and Simon Garrod. 2004b. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Martin J Pickering and Simon Garrod. 2005. Establishing and using routines during dialogue: Implications for psychology and linguistics. *Twenty-first century psycholinguistics: Four cornerstones*.

Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021. A surprisal–duration trade-off across and within the world's languages. In *Proceedings of EMNLP*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Marlou Rasenberg, Asli Özyürek, and Mark Dingemanse. 2020. Alignment in multimodal interaction: An integrative framework. *Cognitive science*, 44(11).

David Reitter, Frank Keller, and Johanna D. Moore. 2006. Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*.

David Reitter and Johanna D. Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76.

David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3).

Arabella Sinclair, Adam Lopez, Christopher G Lucas, and Dragan Gasevic. 2018. Does ability affect alignment in second language tutorial dialogue? In *Proceedings of the 19th annual SIGdial meeting on discourse and dialogue*, pages 41–50.

Arabella J Sinclair, Raquel Fernández, et al. 2021. Construction coordination in first and second language acquisition. In *Proceedings of SemDial*.

Arabella J Sinclair and Bertrand Schneider. 2021. Linguistic and gestural coordination: Do learners converge in collaborative dialogue?. *International Educational Data Mining Society*.

Tanmay Sinha and Justine Cassell. 2015. We click, we align, we learn: Impact of influence and convergence processes on student learning and rapport building. In *Proceedings of the 1st workshop on modeling INTERPERsonal SynchrONy And influence*, pages 13–20.

Anna Siyanova-Chanturia, Kathy Conklin, Sendy Caffarra, Edith Kaan, and Walter JB van Heuven. 2017. Representation and processing of multi-word expressions in the brain. *Brain and language*, 175.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Simone A Sprenger, Willem JM Levelt, and Gerard Kempen. 2006. Lexical access during the production of idiomatic phrases. *Journal of Memory and Language*, 54(2):161–184.

M Tomasello. 2003. Constructing a language: A usage-based theory of language acquisition. *Harvard University Press*.

Antoine Tremblay, Bruce Derwing, Gary Libben, and Chris Westbury. 2011. Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language learning*, 61(2):569–613.

Marten van Schijndel and Tal Linzen. 2018. A neural model of adaptation in reading. In *Proceedings of EMNLP*, Brussels, Belgium.

Arthur Ward and Diane Litman. 2007. Dialog convergence and learning. *Frontiers in Artificial Intelligence and Applications*, 158.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of CogSci*. Cognitive Science Society.

Deanna Wilkes-Gibbs and Herbert H Clark. 1992. Coordinating beliefs in conversation. *Journal of Memory and Language*, 31(2).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, and Morgan et. al Funtowicz. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP: System Demonstrations*.

Yang Xu and David Reitter. 2017. Spectral analysis of information density in dialogue predicts collaborative task performance. In *Proceedings of ACL*.

Yang Xu and David Reitter. 2018. Information density converges in dialogue: Towards an information-theoretic model. *Cognition*, 170.

A. Extraction of Repeated Constructions

After manual analysis of the corpus and extracted constructions used, we define a specific set of vocabulary that is indicative of three broad categories of construction use. We choose to separate constructions referring to *landmarks* from those concerning instructions related to the direction-giving element of the task. The remaining constructions which do not include the vocabulary from either the landmark, or direction lists, we assign to a third, generic category which are more topic independent.

Landmarks The vocabulary for *landmarks* was extracted from the data provided on landmark names within the MapTask corpus, with stop-words removed. This results in the following landmark vocabulary: *mountain, lake, mine, bridge, rope, white,* *diamond, tree, monastery, great, gold, level, slate, disused, stile, trout, farm, telephone, old, collapsed, saxon, barn, dead, rock, shelter, valley, fallen, pillars, outlaws, hideout, field, station, kiosk, river, carved, stones, west, well, east, ruined, rift, fall, truck, banana, viewpoint, indian, country, bandit, territory, stone, saloon, bar, temple, lagoon, totem, pole, broken, gate, granite, quarry, camera, shop, box, flight, museum, pine, rocket, warehouse, roman, baths, extinct, volcano, triangle, secret, start, crevasse, missionary, camp, lost, steps, city, hot, wells, stony, desert, safari, highest, finish, parched, bed, rocks, slabs, noose, attractive, cliffs, remote, village, abandoned, pyramid, water, green, bay, cattle, stockade, wall, circle*.

Directions The vocabulary for *directions* was derived through manual annotation of extracted constructions and contains the following: *go, side, right, left, top, down, bottom, page, move, up, edge, inch, inches, line, centimetres, centimetre, along, across, half, paper, south, map, middle, end, corner, couple, turn, centre, quarters, straight, round, left-hand, above, underneath, right-hand, below, thirds, under, diagonally, towards, halfway, approximately, follow*.

Within *directions*, we observe that there are some ambiguous uses of right: it can be used either as a *generic backchannel*, i.e. "okay right i understand", or as a *direction* "to the right of the". After manually analysing all constructions containing the keyword right, which do not already contain either a landmark or direction keyword (thus already sufficient to differentiate them), we create two categories:
direction right-grams: *right until you 're, right to, the right, its right, right at, your right, my right, your right, to right, going right, right of*
generic right-grams: *right that, right okay, right got it, right you, right i, that 's right, right could*.

A further filtering step is used to correctly recategorise these cases as belonging to either generic or direction. The remaining constructions are categorised as generic.

B. Filled pauses

We define a list of filled pauses according to the part-of-speech tags in MapTask and Switchboard. Following is the filled pauses vocabulary for MapTask : *mmhmm, uh-huh, uh-uh, mm-mm, mm-hmm, mm, erm, eh, ehm, er, um, uh, hmm, nah, huh*. Following is the filled pauses vocabulary for Switchboard : *Hm, Huh, Huh-uh, Uh, Uh-huh, Um, Um-hum, huh, huh-uh, uh, uh-huh, um*.

C. Language Model Fine-Tuning

Using the Hugging Face fine-tuning script, from the *transformers* library (Wolf et al., 2020), we fine-tune GPT-2 (Radford et al., 2019) for 10 epochs, using an early stopping technique to save the best performing model (based on its perplexity).

To obtain estimates for the full Map Task corpus, we iteratively fine-tune GPT-2 on an 80% split of the data, then estimate surprisal over the remaining 20% split, and repeat this for 5 distinct splits. We thus collect surprisal estimates for the full corpus while ensuring that the model has not been fine-tuned on the target dialogues. The pre-trained and fine-tuned perplexity for each split subset is recorded in Table 4.

#	Number of Dialogues	Pre-trained Perplexity	Fine-tuned Perplexity
Set 1	26	191.677	7.282
Set 2	26	208.669	8.874
Set 3	25	210.997	8.450
Set 4	25	202.651	7.792
Set 5	25	197.179	8.261

Table 4: Perplexity of pre-trained and fine-tuned models on the evaluation set.

D. Language Model Adaptation

Learning rate is a critical hyperparameter for ensuring the success of the continuous learning mechanism. An inappropriate learning rate can result in overfitting or underfitting of the language model. To ensure that the model is appropriately tuned, we conduct 18-fold cross-validation tests on six different learning rates: $1e-5$, $1e-4$, $\ldots$, 1. The learning rate is evaluated in terms of the model's change in perplexity after adaptation relative to its perplexity before adaptation. We compute the *adaptation* and *generalisation* performance of each learning rate, and select the one with (i) the highest adaptation performance, and (ii) the smallest difference between adaptation and generalisation performance. Adaptation performance refers to the change in perplexity in the training data (the dialogue at hand) and generalisation performance refers to the change in perplexity in the evaluation data (the remaining 17 dialogues). We refer to the performance on the remaining dialogues as 'generalisation' even though we do not explicitly test for, but only assume, data distribution shifts (cf. Hupkes et al., 2022) . We believe this is a safe assumption to make given the different dialogues are produced from different speakers, talking about different maps, with different landmarks and routes.

E. Non-Construction Sequences

After constructions are extracted from an utterance, the text that remains is made up of non-construction sequences. These sequences can vary in length beyond the range that we find in our constructions. In order to make a fair comparison—in terms of sequence length—between construction and non-construction sequences, we employ a random n-grams sampling method. This approach enables us to examine the information content, facilitating effect, and speech duration of construction and non-construction sequences in this study. We extract all word sequences ranging from 3 to 7 words in length (which are thus the minimum and maximum lengths of constructions) from the entire corpus using this n-gram sampling approach. Of these sequences, we exclude all those which match the construction sequences extracted, those remaining we refer to as *non-construction sequences*. We ensure the length distribution of construction sequences and randomly sampled non-construction sequences match, and ensure there is an equal number of observations for both sequence types. We are then able to compare the attributes of construction sequences with our extracted non-construction sequences.

F. Efficiency

F.1. Facilitating Effect

We investigate how repetition affects facilitating effect through constructing models to predict FE given base (length and position), cumulative, and communicative function factors.

F.1.1. Base Model

```
Formula : logFE10 ~ 1 + logLen + logPositionInTurn

           Mixed Linear Model Regression Results
=========================================================
Model:              MixedLM Dependent Variable: logFE10
No. Observations:   11786   Method:             REML
No. Groups:         252     Scale:              1.3133
Min. group size:    1       Log-Likelihood:     -18452.7283
Max. group size:    305     Converged:          Yes
Mean group size:    46.8
---------------------------------------------------------
                  Coef. Std.Err.   z    P>|z| [0.025 0.975]
---------------------------------------------------------
Intercept         0.958    0.072 13.225 0.000  0.816  1.100
logLen            0.069    0.040  1.749 0.080 -0.008  0.147
logPositionInTurn 0.144    0.043  3.398 0.001  0.061  0.228
Group Var         0.059    0.008
=========================================================
```

F.1.2. Decay & Cumulativity

```
Formula : logFE10 ~ 1 + logLen + logPositionInTurn
                     + logCurrentFreq + logRecencyInTokens

            Mixed Linear Model Regression Results
============================================================
Model:               MixedLM  Dependent Variable:  logFE10
No. Observations:    11751    Method:              REML
No. Groups:          252      Scale:               1.2260
Min. group size:     1        Log-Likelihood:      -17993.8234
Max. group size:     305      Converged:           Yes
Mean group size:     46.6
------------------------------------------------------------
                    Coef.  Std.Err.    z     P>|z| [0.025 0.975]
------------------------------------------------------------
Intercept            1.825    0.092  19.851 0.000  1.645  2.005
logLen               0.173    0.039   4.470 0.000  0.097  0.249
logPositionInTurn  -0.389    0.046  -8.531 0.000 -0.478 -0.300
logCurrentFreq      0.161    0.016  10.272 0.000  0.130  0.192
logRecencyInTokens -0.149    0.006 -26.559 0.000 -0.160 -0.138
Group Var           0.049    0.007
============================================================
```

F.1.3. Communicative Function

```
Formula : logFE10 ~ 1 + logLen + logPositionInTurn
                      + logCurrentFreq + logRecencyInTokens
                      + ConstructionType

                  Mixed Linear Model Regression Results
=========================================================================
Model:                     MixedLM       Dependent Variable:       logFE10
No. Observations:          11751         Method:                   REML
No. Groups:                252           Scale:                    1.1793
Min. group size:           1             Log-Likelihood:           -17772.0695
Max. group size:           305           Converged:                Yes
Mean group size:           46.6
-------------------------------------------------------------------------
                                Coef.  Std.Err.    z    P>|z| [0.025 0.975]
-------------------------------------------------------------------------
Intercept                        1.232    0.094  13.066 0.000  1.048  1.417
ConstructionType[T.Landmark]     0.634    0.030  21.464 0.000  0.576  0.691
ConstructionType[T.Direction]    0.195    0.024   8.117 0.000  0.148  0.242
logLen                           0.175    0.038   4.584 0.000  0.100  0.250
logPositionInTurn               -0.265    0.045  -5.869 0.000 -0.353 -0.176
logCurrentFreq                   0.207    0.016  13.305 0.000  0.177  0.238
logRecencyInTokens              -0.116    0.006 -20.154 0.000 -0.127 -0.104
Group Var                        0.049    0.007
=========================================================================
```

F.2. Speech Rate - Mean Token Duration

To explore the effects on the rate of delivery, we construct models with base factors (construction length in tokens, number of characters per word, and average word frequency obtained from a corpus of subtitles by Brysbaert et al. (2012)). We then add Decay and Cumulativity, Communicative Function and Information factors.

F.2.1. Base

```
Formula : logTokenMeanDuration ~ 1 + logLen + logMeanCharacter
                                   + logMeanTokenFrequency

              Mixed Linear Model Regression Results
===================================================================
Model:              MixedLM Dependent Variable: logTokenMeanDuration
No. Observations:   11786   Method:             REML
No. Groups:         252     Scale:              0.1775
Min. group size:    1       Log-Likelihood:     -6777.4806
Max. group size:    305     Converged:          Yes
Mean group size:    46.8
-------------------------------------------------------------------
                        Coef.  Std.Err.    z    P>|z| [0.025 0.975]
-------------------------------------------------------------------
Intercept               -1.585    0.086 -18.413 0.000 -1.754 -1.416
logLen                  -0.023    0.015  -1.559 0.119 -0.052  0.006
logMeanCharacter         0.685    0.016  44.049 0.000  0.655  0.716
logMeanTokenFrequency   -0.965    0.026 -36.629 0.000 -1.016 -0.913
Group Var                0.032    0.009
===================================================================
```

F.2.2. Decay & Cumulativity

```
Formula : logTokenMeanDuration ~ 1 + logLen + logMeanCharacter
                                   + logMeanTokenFrequency
                                   + logCurrentFreq
                                   + logRecencyInTokens

              Mixed Linear Model Regression Results
===================================================================
Model:              MixedLM Dependent Variable: logTokenMeanDuration
No. Observations:   11751   Method:             REML
No. Groups:         252     Scale:              0.1750
Min. group size:    1       Log-Likelihood:     -6684.8476
Max. group size:    305     Converged:          Yes
Mean group size:    46.6
-------------------------------------------------------------------
                        Coef.  Std.Err.    z    P>|z| [0.025 0.975]
-------------------------------------------------------------------
Intercept               -1.589    0.086 -18.405 0.000 -1.758 -1.419
logLen                  -0.052    0.015  -3.531 0.000 -0.082 -0.023
logMeanCharacter         0.693    0.015  44.758 0.000  0.663  0.723
logMeanTokenFrequency   -0.949    0.027 -35.759 0.000 -1.001 -0.897
logCurrentFreq          -0.061    0.006 -10.124 0.000 -0.073 -0.049
logRecencyInTokens       0.012    0.002   6.209 0.000  0.008  0.016
Group Var                0.032    0.009
===================================================================
```

F.2.3. Communicative Function

```
Formula : logTokenMeanDuration ~ 1 + logLen + logMeanCharacter
                                 + logMeanTokenFrequency
                                 + logCurrentFreq
                                 + logRecencyInTokens
                                 + ConstructionType
```

Mixed Linear Model Regression Results

Model:	MixedLM	Dependent Variable:	logTokenMeanDuration
No. Observations:	11751	Method:	REML
No. Groups:	252	Scale:	0.1520
Min. group size:	1	Log-Likelihood:	-5866.3916
Max. group size:	305	Converged:	Yes
Mean group size:	46.6		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-2.388	0.092	-25.993	0.000	-2.568	-2.208
ConstructionType[T.Landmark]	0.360	0.016	22.391	0.000	0.328	0.391
ConstructionType[T.Direction]	0.410	0.010	41.979	0.000	0.391	0.429
logLen	-0.144	0.014	-10.236	0.000	-0.171	-0.116
logMeanCharacter	0.625	0.015	42.987	0.000	0.596	0.653
logMeanTokenFrequency	-0.606	0.030	-19.947	0.000	-0.666	-0.547
logCurrentFreq	-0.046	0.006	-8.135	0.000	-0.057	-0.035
logRecencyInTokens	0.012	0.002	6.603	0.000	0.009	0.016
Group Var	0.029	0.008				

F.2.4. Information Factors

```
Formula : logTokenMeanDuration ~ 1 + logLen + logMeanCharacter
                                 + logMeanTokenFrequency
                                 + S + logFE10
```

Mixed Linear Model Regression Results

Model:	MixedLM	Dependent Variable:	logTokenMeanDuration
No. Observations:	11786	Method:	REML
No. Groups:	252	Scale:	0.1756
Min. group size:	1	Log-Likelihood:	-6722.4105
Max. group size:	305	Converged:	Yes
Mean group size:	46.8		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-1.735	0.088	-19.629	0.000	-1.908	-1.562
logLen	-0.004	0.015	-0.260	0.795	-0.033	0.025
logMeanCharacter	0.689	0.015	44.480	0.000	0.658	0.719
logMeanTokenFrequency	-0.971	0.026	-36.818	0.000	-1.023	-0.920
S	0.045	0.004	10.245	0.000	0.037	0.054
logFE10	0.021	0.005	4.146	0.000	0.011	0.031
Group Var	0.031	0.009				

F.2.5. Full

```
Formula : logTokenMeanDuration ~ 1 + logLen + logMeanCharacter
                                 + logMeanTokenFrequency
                                 + logCurrentFreq
                                 + logRecencyInTokens
                                 + ConstructionType
                                 + S + logFE10
```

Mixed Linear Model Regression Results

Model:	MixedLM	Dependent Variable:	logTokenMeanDuration
No. Observations:	11751	Method:	REML
No. Groups:	252	Scale:	0.1502
Min. group size:	1	Log-Likelihood:	-5807.5897
Max. group size:	305	Converged:	Yes
Mean group size:	46.6		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-2.549	0.093	-27.375	0.000	-2.732	-2.367
ConstructionType[T.Landmark]	0.372	0.016	23.059	0.000	0.340	0.403
ConstructionType[T.Direction]	0.416	0.010	42.741	0.000	0.397	0.435
logLen	-0.122	0.014	-8.643	0.000	-0.149	-0.094
logMeanCharacter	0.625	0.014	43.257	0.000	0.597	0.654
logMeanTokenFrequency	-0.599	0.030	-19.825	0.000	-0.659	-0.540
logCurrentFreq	-0.037	0.006	-6.532	0.000	-0.048	-0.026
logRecencyInTokens	0.008	0.002	4.291	0.000	0.004	0.012
S	0.044	0.004	10.592	0.000	0.036	0.052
logFE10	0.019	0.005	4.030	0.000	0.010	0.028
Group Var	0.029	0.008				

G. Effectiveness - Predicting Task Success

G.1. Base

Duration, Information and Construction Usage
We consider dialogue length and duration, proportion of construction usage, then factors specific to construction usage: length, frequency (cumulativity), repetition distance (recency), normalised count of unique constructions, construction PMI. We also include information measures: global utterance and construction information rate, construction FE, and how it changes with repetition, global uniformity of information, and within utterance uniformity.

OLS Regression Results

Dep. Variable:	log_success	R-squared:	0.041
Model:	OLS	Adj. R-squared:	0.034
Method:	Least Squares	F-statistic:	5.423
		Prob (F-statistic):	0.0215
		Log-Likelihood:	-142.58
No. Observations:	128	AIC:	289.2
Df Residuals:	126	BIC:	294.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.4715	0.202	-22.105	0.000	-4.872	-4.071
log_total_dialogue_token	0.9022	0.387	2.329	0.021	0.136	1.669

Omnibus:	7.860	Durbin-Watson:	1.743
Prob(Omnibus):	0.020	Jarque-Bera (JB):	7.524
Skew:	0.540	Prob(JB):	0.0232
Kurtosis:	3.495	Cond. No.	7.37

G.2. Communicative Function

We split factors in our base model that relate to constructions (length, frequency (cumulativity), repetition distance (recency), normalised count of unique constructions, construction PMI, surprisal, FE and their slope) by communicative function: Generic, Direction and Landmark.

OLS Regression Results

Dep. Variable:	log_success	R-squared:	0.173
Model:	OLS	Adj. R-squared:	0.137
Method:	Least Squares	F-statistic:	4.824
		Prob (F-statistic):	0.000485
		Log-Likelihood:	-126.41
No. Observations:	121	AIC:	264.8
Df Residuals:	115	BIC:	281.6
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-5.5190	0.490	-11.255	0.000	-6.490	-4.548
log_total_dialogue_token	1.3038	0.422	3.093	0.002	0.469	2.139
landmark_constr_surprisal	-1.4415	0.417	-3.460	0.001	-2.267	-0.616
generic_constr_surprisal	1.5570	0.632	2.463	0.015	0.305	2.809
generic_constr_fe	1.6789	0.663	2.533	0.013	0.366	2.992
direction_constr_fe_slope	1.5217	0.757	2.009	0.047	0.021	3.022

Omnibus:	5.996	Durbin-Watson:	1.527
Prob(Omnibus):	0.050	Jarque-Bera (JB):	5.686
Skew:	0.411	Prob(JB):	0.0582
Kurtosis:	3.672	Cond. No.	19.2

G.3. Shared Vs. Unshared Constructions

We then investigate whether the same set of factors (including communicative function) , but differentiating between whether the constructions considered are shared (repeated by both speakers) or unshared (repeated only by one speaker).

G.3.1. Unshared Constructions: Self repetition

OLS Regression Results

Dep. Variable:	log_success	R-squared:	0.327
Model:	OLS	Adj. R-squared:	0.285
Method:	Least Squares	F-statistic:	7.704
		Prob (F-statistic):	1.43e-07
		Log-Likelihood:	-111.77
No. Observations:	119	AIC:	239.5
Df Residuals:	111	BIC:	261.8
Df Model:	7		
Covariance Type:	nonrobust		

```
                              coef  std err      t   P>|t|   [0.025 0.975]
---------------------------------------------------------------------------
Intercept                  -4.5066  0.498  -9.055  0.000  -5.493  -3.520
landmark_constr_length      1.4124  0.359   3.929  0.000   0.700   2.125
landmark_constr_freq       -1.3821  0.566  -2.440  0.016  -2.50   -0.260
direction_constr_pmi       -0.7079  0.285  -2.483  0.015  -1.273  -0.143
landmark_constr_s          -1.0803  0.400  -2.701  0.008  -1.873  -0.288
generic_constr_s_slope      1.4395  0.587   2.452  0.016   0.27    2.603
direction_constr_fe_slope   2.1167  0.566   3.739  0.000   0.995   3.239
token_s_variance            0.7319  0.353   2.073  0.040   0.032   1.431
==========================================================================
Omnibus:                         5.193   Durbin-Watson:              1.788
Prob(Omnibus):                   0.075   Jarque-Bera (JB):           5.861
Skew:                            0.252   Prob(JB):                  0.0534
Kurtosis:                        3.963   Cond. No.                    18.4
==========================================================================
```

G.3.2. Shared Constructions: Between-speaker repetition

```
==========================================================================
                            OLS Regression Results
==========================================================================
Dep. Variable:              log_success   R-squared:                 0.574
Model:                              OLS   Adj. R-squared:            0.406
Method:                   Least Squares   F-statistic:               3.413
                                          Prob (F-statistic):      0.00110
                                          Log-Likelihood:          -42.782
No. Observations:                    54   AIC:                       117.6
Df Residuals:                        38   BIC:                       149.4
Df Model:                            15
Covariance Type:              nonrobust
==========================================================================
                              coef   std err      t   P>|t|  [0.025  0.975]
---------------------------------------------------------------------------
Intercept                  -5.6640  1.025 -5.528   0.000  -7.738  -3.590
constr_proportion           2.5749  0.789  3.265   0.002   0.978   4.172
landmark_constr_length     -1.1985  0.577 -2.079   0.044  -2.366  -0.031
direction_constr_length     3.0891  1.373  2.250   0.030   0.309   5.869
generic_constr_frequency    1.9821  0.932  2.127   0.040   0.096   3.868
direction_constr_rep_dist   2.5147  0.777  3.237   0.003   0.942   4.087
generic_constr_rep_dist     2.2857  0.898  2.545   0.015   0.467   4.104
landmark_constr_inventory  -1.9448  0.610 -3.189   0.003  -3.179  -0.710
direction_constr_inventory -2.2311  0.959 -2.325   0.02   -4.173  -0.289
landmark_constr_s          -6.1818  1.444 -4.280   0.000  -9.106  -3.258
direction_constr_s          2.1291  0.764  2.785   0.008   0.582   3.677
landmark_constr_s_slope     2.2769  1.023  2.226   0.032   0.207   4.347
generic_constr_s_slope      1.8761  0.688  2.726   0.010   0.483   3.269
landmark_constr_fe         -2.3585  0.979 -2.410   0.021  -4.340  -0.377
generic_constr_fe          -1.3270  0.591 -2.244   0.031  -2.524  -0.130
token_s_variance            1.6259  0.673  2.417   0.021   0.264   2.988
==========================================================================
Omnibus:                         0.938   Durbin-Watson:              1.916
Prob(Omnibus):                   0.626   Jarque-Bera (JB):           0.322
Skew:                            0.069   Prob(JB):                   0.851
Kurtosis:                        3.352   Cond. No.                    35.9
==========================================================================
```

G.4. Speaker Role

We then run the same model as in Communicative Function, with factors split by speaker role (Giver: *g*, Follower: *f*).

```
==========================================================================
                            OLS Regression Results
==========================================================================
Dep. Variable:              log_success   R-squared:                 0.352
Model:                              OLS   Adj. R-squared:            0.212
Method:                   Least Squares   F-statistic:               2.521
                                          Prob (F-statistic):      0.00619
                                          Log-Likelihood:          -70.010
No. Observations:                    80   AIC:                       170.0
Df Residuals:                        65   BIC:                       205.7
Df Model:                            14
Covariance Type:              nonrobust
==========================================================================
                               coef  std err     t   P>|t|  [0.025  0.975]
---------------------------------------------------------------------------
Intercept                    -6.9228  1.152 -6.007  0.000  -9.224  -4.621
generic_constr_rep_dist_g     1.8630  0.577  3.227  0.002   0.710   3.016
generic_constr_rep_dist_f    -1.2466  0.533 -2.339  0.022  -2.311  -0.182
direction_constr_rep_freq_g  -0.9199  0.451 -2.042  0.045  -1.820  -0.020
direction_constr_rep_freq_f   2.3764  1.004  2.367  0.021   0.372   4.381
generic_constr_length_g       1.7571  0.776  2.264  0.027   0.207   3.307
landmark_constr_inventory_g   1.0153  0.421  2.409  0.019   0.173   1.857
landmark_constr_inventory_f  -1.2301  0.462 -2.665  0.010  -2.152  -0.308
landmark_constr_pmi_f        -1.6218  0.448 -3.619  0.001  -2.517  -0.727
direction_constr_s_f         -1.9093  0.844 -2.263  0.027  -3.595  -0.224
generic_constr_s_g            1.8096  0.618  2.927  0.005   0.575   3.045
direction_constr_fe_f        -2.6597  0.840 -3.167  0.002  -4.337  -0.983
generic_constr_fe_f           1.7736  0.591  3.003  0.004   0.594   2.953
landmark_constr_fe_slope_g    2.2363  0.796  2.810  0.007   0.647   3.826
generic_constr_fe_slope_g     7.3357  2.433  3.016  0.004   2.477  12.194
==========================================================================
Omnibus:                         1.354   Durbin-Watson:              1.643
Prob(Omnibus):                   0.508   Jarque-Bera (JB):           1.290
Skew:                            0.186   Prob(JB):                   0.525
Kurtosis:                        2.501   Cond. No.                    68.2
==========================================================================
```