Construction Repetition Reduces Information Rate in Dialogue

Giulianelli, M.; Sinclair, A.; Fernández, R.

Publication date
2022
Document Version
Final published version
Published in
The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing

Construction Repetition Reduces Information Rate in Dialogue

Mario Giulianelli
Institute for Logic, Language and Computation
University of Amsterdam
m.giulianelli@uva.nl

Arabella Sinclair
Department of Computing Science
University of Aberdeen
arabella.sinclair@abdn.ac.uk

Raquel Fernández
Institute for Logic, Language and Computation
University of Amsterdam
raquel.fernandez@uva.nl

Abstract

Speakers repeat constructions frequently in dialogue. Due to their peculiar information-theoretic properties, repetitions can be thought of as a strategy for cost-effective communication. In this study, we focus on the repetition of lexicalised constructions—i.e., recurring multi-word units—in English open-domain spoken dialogues. We hypothesise that speakers use *construction repetition* to mitigate information rate, leading to an overall decrease in utterance information content over the course of a dialogue. We conduct a quantitative analysis, measuring the information content of constructions and that of their containing utterances, estimating information content with an adaptive neural language model. We observe that construction usage lowers the information content of utterances. This *facilitating effect* (i) increases throughout dialogues, (ii) is boosted by repetition, (iii) grows as a function of repetition frequency and density, and (iv) is stronger for repetitions of referential constructions.

1 Introduction

The repeated use of particular configurations of structures and lexemes, *constructions*, is pervasive in conversational language use (Tomasello, 2003; Goldberg, 2006). Such repetition can be understood as a surface level signal of processes of coordination (Sinclair and Fernández, 2021) or 'interpersonal synergy' between conversational partners (Fusaroli et al., 2014). Speakers may use repetitions to successfully maintain common ground with their interlocutors (Brennan and Clark, 1996; Pickering and Garrod, 2004), because they are primed by their recent linguistic experience (Bock, 1986), or to avoid a costly on-the-fly search for alternative phrasings (see, e.g., Kuiper, 1995). At the same time, repetitions are also advantageous for comprehenders. Repeating a sequence of words positively reshapes expectations for those words, allowing comprehenders to process them more rapidly (for a review, see Bigand et al., 2005). As speakers are known to take into consideration both their own production cost and their addressee's processing effort (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989; Frank and Goodman, 2012), its two-sided processing advantage, as described above, makes construction repetition an efficient, cost-reducing communication strategy. In this paper, we investigate whether and how these information-theoretic properties of repetitions shape patterns of information rate in open-domain spoken dialogue.

Information theory is the study of the conditions affecting the transmission and processing of information. To the foundations of the field belongs the noisy-channel coding theorem (Shannon, 1948), which states that for any given degree of noise in a communication channel, it is possible to communicate discrete signals nearly error-free up to a maximum information rate, the *channel capacity*. If speakers use the communication channel optimally, they might send information at a rate that is always close to the channel capacity. This observation is at the basis of the principle of Entropy Rate Constancy (ERC; Genzel and Charniak, 2002), which predicts that the information rate of speaker's utterances, measured as the utterance conditional entropy (i.e., its in-context *Shannon information content* or *information density*) remains constant throughout discourse. The ERC predictions have been empirically confirmed for written language production (Genzel and Charniak, 2002, 2003; Qian and Jaeger, 2011) but results on dialogue are mixed (Vega and Ward, 2009; Doyle and Frank, 2015b,a; Xu and Reitter, 2018; Giulianelli et al., 2021), with some studies suggesting a decreasing information rate over the

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 665–682
November 20–23, 2022.

course of dialogues (Vega and Ward, 2009; Giulianelli and Fernández, 2021). We hypothesise that this decreasing trend in dialogue may be associated with construction repetition. We conjecture that speakers use construction repetition as a strategy for information rate mitigation, by padding the more information dense parts of their utterances with progressively less information dense constructions—leading to an overall decrease in information rate over the course of a dialogue.

We extract occurrences of fully lexicalised constructions (see Table 1 for examples) from a corpus of open-domain spoken dialogues and use a Transformer-based neural language model to estimate their contribution to utterance information content. First, we confirm that constructions indeed exhibit lower information content than other expressions and that information content further decreases when constructions are repeated. Then, we show that the decreasing trend of information content observed *over utterances*—which contradicts the ERC principle—is driven by the increasing mitigating effect of construction repetition, measured as a construction's (increasingly) negative contribution to the information content of its containing utterance, what we call its *facilitating effect*.

In sum, our study provides new empirical evidence that dialogue partners use construction repetition as a strategy for information rate mitigation, which can explain why the rate of information transmission in dialogue, in contrast to the constancy predicted by the theory (Genzel and Charniak, 2002), is often found to decrease. Our findings inform the development of better dialogue models. They indicate, as suggested in related work (e.g., Xi et al., 2021), that while avoiding *degenerate* repetitions in utterance generation (Li et al., 2016; Welleck et al., 2019) is an appropriate strategy, dialogue systems should not suppress *human-like* patterns of repetition as these make automatic systems be perceived as more natural and more effective in conversational settings.

SPXV	**SAXQ**	**S9YG**
want to be with him	*it on the television*	I bet you can
shit like that	*for a family*	yeah I used to
I can be	think that's a	*go to bed*
to see her	*the orient express*	and I love
and she just	one thing that	*the window and*
I quite like	*one of my favourites*	and I think it's
you don't like	*on the television*	yeah I think so
and you're like	yes yeah I	*the same people*
going to go	erm I think	is she in
you're going to	a really good	*lock the door*

Table 1: Top 10 constructions from three dialogues of the Spoken BNC (Love et al., 2017), sorted according to the PMI between a construction and its dialogue (§6.1). Referential constructions in italics (§3.1). Headers correspond to the dialogues' IDs in the corpus.

2 Background

2.1 Constructions

This work focuses on *constructions*, seen as particular configurations of structures and lexemes in usage-based accounts of natural language (Tomasello, 2003; Bybee, 2006, 2010; Goldberg, 2006). According to these accounts, models of language processing must consider not only individual lexical elements according to their syntactic roles but also more complex form-function units, which can break regular phrasal structures—e.g., *'I know I'*, *'something out of'*. We further focus on fully lexicalised constructions (sometimes called *formulaic expressions*, or *multi-word expressions*). Commonly studied types of constructions are idioms (*'break the ice'*), collocations (*'pay attention to'*), phrasal verbs (*'make up'*), and lexical bundles (*'a lot of the'*). In §3.1, we explain how the notion of lexicalised construction is operationalised in the current study; Table 1 shows some examples.

A common property of constructions is their frequent occurrence in natural language. As such, they possess what, in usage-based accounts, is sometimes referred to as 'processing advantage' (Conklin and Schmitt, 2012; Carrol and Conklin, 2020). Evidence for the processing advantage of construction usage has been found in reading (Arnon and Snider, 2010; Tremblay et al., 2011), naming latency (Bannard and Matthews, 2008; Janssen and Barber, 2012), eye-tracking (Underwood et al., 2004; Siyanova-Chanturia et al., 2011), and electrophysiology (Tremblay and Baayen, 2010; Siyanova-Chanturia et al., 2017). In this paper, we model this processing advantage as reduced information content and show that it can mitigate information rate throughout entire dialogues.

2.2 Information Content, Surprisal, and Processing Effort

Estimates of information content have been shown to be good predictors of processing effort in perception (Jelinek et al., 1975; Clayards et al., 2008), reading (Keller, 2004; Demberg and Keller, 2008; Levy et al., 2009), and sentence interpretation

(Levy, 2008; Gibson et al., 2013). In these studies, information content is typically referred to as *surprisal*, taken as a measure of how unpredictable, unlikely, or surprising a linguistic signal is in its context. As speakers take into consideration their addressee's processing effort (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989), their linguistic choices can often be explained as strategies to manage the fluctuations of information content over time. Surprisal-based accounts have indeed been successful at explaining various aspects of language production: speakers tend to reduce the duration of less surprising sounds (Aylett and Turk, 2004, 2006; Bell et al., 2003; Demberg et al., 2012); they are more likely to drop sentential material within less surprising scenarios (Jaeger and Levy, 2007; Frank and Jaeger, 2008; Jaeger, 2010); they tend to overlap at low-surprisal dialogue turn transitions (Dethlefs et al., 2016); and they produce sentences at a constant information rate in texts (Genzel and Charniak, 2002; Qian and Jaeger, 2011; Giulianelli and Fernández, 2021).

To measure information content we use GPT-2 (Radford et al., 2019), a neural language model. We thereby follow the established approach (e.g., Genzel and Charniak, 2002; Keller, 2004; Xu and Reitter, 2018) of using language models to estimate information content. Neural models' estimates in particular have been shown to be good predictors of processing effort, measured as reading time, gaze duration, and N400 response (Monsalve et al., 2012; Goodkind and Bicknell, 2018; Merkx and Frank, 2021; Schrimpf et al., 2021). We further implement a simple neural adaptation mechanism, performing continuous gradient updates based on utterance prediction error; this not only leads to a more psychologically plausible model but also to the estimation of more human-like expectations (van Schijndel and Linzen, 2018).

3 Data

We conduct our study on the Spoken British National Corpus[1] (Love et al., 2017), a dataset of transcribed open-domain spoken dialogues containing 1,251 contemporary British English conversations, collected in a range of real-life contexts. We focus on the 622 dialogues that feature only two speakers, and randomly split them into a 70% finetuning set (to be used as described in §4) and a 30% analysis set (used in our experiments, as described in §5 and §6). Table 2 shows some statistics of the dialogues used in this study.

	Mean ± Sd	Median	Min	Max
Dialogue length (# utterances)	736 ± 599	541.5	67	4859
Dialogue length (# words)	7753 ± 5596	6102	819	39575
Utterance length (# words)	11 ± 15	6	1	982

Table 2: Two-speaker dialogue statistics, Spoken BNC.

3.1 Extracting Repeated Constructions

We define constructions as multi-word sequences repeated within a dialogue. To extract constructions from each dialogue, we use the sequential pattern mining method proposed by Duplessis et al. (2017a,b, 2021), which treats the extraction task as an instance of the longest common subsequence problem (Hirschberg, 1977; Bergroth et al., 2000).[2] We modify it to not discard multiple repetitions of a construction that occur in the same utterance. We focus on constructions of at least three tokens, uttered at least three times in a dialogue by any of the dialogue participants. Repeated sequences that mostly appear as a sub-part of a larger construction are discarded.[3] We also exclude sequences containing punctuation marks or which consist of more than 50% filled pauses (e.g., *'mm'*, *'erm'*).[4]

Applying the described extraction procedure to the 187 dialogues in the analysis split of the Spoken BNC yields a total of 5,893 unique constructions and 60,494 occurrences. Further statistics of the extracted constructions are presented in Table 3, and Table 1 shows 10 example constructions extracted from three dialogues. For analysis purposes, we distinguish between referential and non-referential constructions. We label a construction as *referential* if it includes nouns, unless the nouns are highly generic.[5] Referential constructions are mostly topic-determined; examples are *'playing table tennis'*, *'a woolly jumper'*, *'a room with a view'*. The remaining constructions are labelled as *non-referential*. These mainly include topic-independent expressions and conversational markers, such *'a lot of'*, *'I don't know'*, and *'yes of course'*. Our dataset consists of 5,291 referen-

[1] http://www.natcorp.ox.ac.uk.

[2] Their code is freely available at https://github.com/GuillaumeDD/dialign.

[3] We discard constructions that appear less than twice outside of a larger repeated construction in a given dialogue (e.g., *'think of it'* vs. *'think of it like'*).

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

[5] We define a limited specific vocabulary of generic nouns (e.g., *'thing'*, *'fact'*, *'time'*); full vocabulary in Appendix B.

tial and 55,203 non-referential construction occurrences, 1,143 and 4,750 construction forms; see Table 1 for further examples.

	Mean ± Sd	Median	Max
Construction Length	3.27 ± 0.58	3	7
Construction Frequency	4.29 ± 3.04	3	70
Constructions per Dialogue	325.34 ± 458.64	149	2817
Referential	30.96 ± 39.75	19	346
Non-Referential	296.88 ± 424.17	134.5	2530
Utterance Length	31.19 ± 36.19	21	959

Table 3: Construction statistics for our analysis split of the Spoken BNC. *Constr. Frequency*: occurrences of a given construction in a dialogue. *Constr. per Dialogue*: occurrences of all constructions in a dialogue. *Utterance Length*: number of words in utterances containing a construction. The minimum is always 3 by design (§3.1). The difference between referential and non-referential is only significant for *Constr. per Dialogue*.

4 Experimental Setup

In this section, we define our information-theoretic measures and present the adaptive language model used to produce information content estimates.[6]

4.1 Information Content Measures

The *information content* of a word choice w_i is the negative logarithm of the corresponding word probability, conditioned on the utterance context $u_{:w_i}$ (i.e., the words that precede w_i in utterance u) and on the local dialogue context l:

$$H(w_i|u_{:w_i}, l) = -\log_2 P(w_i|u_{:w_i}, l) \quad [1]$$

We define the local dialogue context l as the 50 tokens that precede the first word in the utterance.[7] We use tokens as a unit of context size, rather than utterances, since they more closely correspond to the temporal units used in previous work (e.g., Reitter et al., 2006), and since the length of utterances can vary significantly (see Table 2). To measure the information content of a construction c, we average over word-level information content values:

$$H(c; u_{:c}, l) = \frac{1}{|c|} \sum_{w_i \in c} H(w_i|u_{:c}, l) \quad [2]$$

[6]Code and statistical analysis are available at https://github.com/dmg-illc/uid-dialogue.

[7]Building on prior work (Reitter et al., 2006) that uses a window of 15 seconds of spoken dialogue as the locus of local repetition effects, we compute the average speech rate in the Spoken BNC (3.16 tokens/second) and multiply it by 15; we then round up the result (47.4) to 50 tokens.

We use the same averaging strategy to compute the information content of entire utterances, following prior work (e.g., Genzel and Charniak, 2002; Xu and Reitter, 2018):

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

The above information content estimates target constructions and entire utterances but they do not qualify the relationship between the two. We also measure the information content change (increase or reduction in information rate) contributed by a construction c to its containing utterance, which we call the *facilitating effect* of a construction. Facilitating effect is defined as the logarithm of the ratio between the information content of a construction and that of its utterance context:

$$FE(c; u, l) = \log_2 \frac{\frac{1}{|u|-|c|} \sum_{c \not\ni w_j \in u} H(w_j|u_{:w_i}, l)}{\frac{1}{|c|} \sum_{w_i \in c} H(w_i|u_{:c}, l)} \quad [4]$$

By definition, this quantity is positive when the construction has lower information content than its context, and negative when it has higher information content. When the utterance consists of a single construction, facilitating effect is set to 0.

We can expect the values produced by our information content and facilitating effect measurements (Eq. 2 and 4, respectively) to correlate: it is more likely for a construction to have a (positive) facilitating effect if its information content is low. When a construction's information content is high, the information content of its utterance context must be even greater for facilitating effect to occur. Nevertheless, perfect correlation does not follow a priori from the definition of the two measures; we will show this empirically in §5.4.

4.2 Language Model

To estimate the per-word conditional probabilities that are necessary to compute information content (Eq. 1), we use an adaptive language model. The model is conditioned on local contextual cues via an attention mechanism (Vaswani et al., 2017) and it learns continually (see, e.g., Krause et al., 2018) from exposure to the global dialogue context. We use GPT-2 (Radford et al., 2019), a pre-trained autoregressive Transformer language model. We rely on HuggingFace's implementation of GPT-2 with default tokenizers and parameters (Wolf et al., 2020) and finetune the pre-trained model on a 70%

training split of the Spoken BNC to adapt it to the idiosyncrasies of spoken dialogic data.[8] We refer to this finetuned version as the *frozen* model. We use an attention window of length $|u_{:w_i}| + 50$, i.e., the sum of the utterance length up to word w_i and the size of the local dialogue context.

As a continual learning mechanism, we use back-propagation on the cross-entropy next word prediction error, a simple yet effective adaptation approach motivated in §2.2. Following van Schijndel and Linzen (2018), when estimating information content for a dialogue, we begin by processing the first utterance using the frozen language model and then gradually update the model parameters after each turn. For these updates to have the desired effect, the learning rate should be appropriately tuned. It should be sufficiently high for the language model to adapt during a single dialogue, yet an excessively high learning rate can cause the language model to lose its ability to generalise across dialogues. To find the appropriate rate, we randomly select 18 dialogues from the analysis split of the Spoken BNC[9] and run an 18-fold cross-validation for a set of six candidate learning rates: $1\mathrm{e}-5$, $1\mathrm{e}-4$, ..., 1. We finetune the model on each dialogue using one of these learning rates and compute perplexity reduction (i) on the dialogue itself (*adaptation*) as well as (ii) on the remaining 17 dialogues (*generalisation*). We select the learning rate yielding the best adaptation over cross-validation folds ($1\mathrm{e}-3$), while still improving the model's generalisation ability. See Appendix C.2 for further details.

5 Preliminary Experiments

In this section, we present preliminary experiments on the information content of utterances and constructions, which set the stage for our analysis of the facilitating effect of construction repetition.

5.1 Utterance Information Content

Our experiments are motivated by the mixed results on the dynamics of information rate in dialogue discussed in §1. We thus begin by testing if the Entropy Rate Constancy (ERC) principle holds in the Spoken BNC, i.e., whether utterance information content remains stable over the course of a dialogue. Following a procedure established in prior work (Xu and Reitter, 2018), we fit a linear mixed effect model with the logarithm of utterance position and construction length as fixed effects (we will refer to their coefficients as β), and include multi-level random effects grouped by dialogue. *For the ERC principle to hold, the position of an utterance within a dialogue should have no effect on its information content.* Instead, we find that utterance information content decreases significantly over time ($\beta = -0.119, p < 0.005$, *95% c.i.* $-0.130 : -0.108$), in line with previous negative results on open-domain and task-oriented dialogue (Vega and Ward, 2009; Giulianelli and Fernández, 2021). The strongest drop occurs in the first ten dialogue utterances ($\beta = -0.886, p < 0.005$, *95% c.i.* $-0.954 : -0.818$) but the decrease is still significant for later utterances ($\beta = -0.043, p < 0.005$, *95% c.i.* $-0.054 : -0.032$).

5.2 Construction Information Content

Our hypothesis that construction repetition progressively reduces the information rate of utterances is motivated by the fact that constructions are known to have a processing advantage (see §1 and §2.1). This property makes them an efficient production strategy, i.e., one that reduces the speaker's and addressee's collaborative effort. Before investigating if the hypothesised information rate mitigation strategy is at play, we test whether our information theoretic measures and the language model used to generate them are able to capture processing advantage: *we expect our framework to yield lower information content estimates (Eq. 2) for constructions than for other word sequences.* Indeed, the information content of constructions is significantly lower than that of non-construction sequences ($t = -168.82, p < 0.005$, *95% c.i.* $-2.033 : -1.987$).[10] Constructions' information content is on average 2 bits lower than that of non-constructions. We conclude that our estimates of information content are a sensible model of the processing advantage of constructions.

5.3 Stable Rate of Construction Usage

In experiment §5.2, we confirmed that constructions have lower information content than other utterance material. A simple strategy to decrease

[8] More details on finetuning can be found in Appendix C.1.

[9] This amounts to ca. 10% of the analysis split. We use the analysis split because there is no risk of "overfitting" with respect to our main analyses.

[10] We extract all 3- to 7-grams from our analysis split of the Spoken BNC, excluding all n-grams that are equal to extracted constructions. We then sample, for each length n from 3 to 7, s_n non-construction sequence occurrences—where s_n is the number of occurrences of n-tokens-long extracted constructions.. The length distributions should match because length has an effect on *S* and *FE* (see §6.3).

utterance information content over dialogues (we do observe this decrease in the Spoken BNC, as described in §5.1) could then simply be to increase the rate of construction usage. To test if this strategy is at play, we fit a linear mixed effect model with utterance position as the predictor and the proportion of construction tokens in an utterance as the response variable. Over the course of a dialogue, the increase in the proportion of an utterance's tokens which belong to a construction is negligible ($\beta = 0.004, p < 0.05$, *95% c.i.* $0.001 : 0.008$). Speakers produce constructions at a stable rate (see also Figure 2 in Appendix B), indicating that an alternative strategy for information rate reduction is at work.

5.4 Information Content vs. Facilitating Effect

The facilitating effect *FE* of a construction is a function of its information content and the information content of its containing utterance (Eq. 4). To ensure that our estimates of *FE* are not entirely determined by construction information content (cf. §4.1), we inspect the relation between the two measures empirically, by looking at the values they take in our dataset of constructions. We find that the Kendall's rank-correlation between *FE* and information content is -0.623 ($p < 0.005$): although this is a rather strong negative correlation, the fact that the score is not closer to -1 indicates that there are cases where the two values are both either high or low. We indeed find examples of constructions with high information content *H* and high facilitating effect *FE*:

A: we'll level that right press p purchase and
B: right
A: **go back to** recommended ***(H = 5.30 FE = 1.65)***

as well cases where information content is low and facilitating effect is low or negative:

A: right let's go and have a drink
B yeah
A: **let's go and have** a drink ***(H = 2.10 FE = −2.21)***

These examples have been selected among occurrences with *H*/*FE* higher or lower than the mean *H*/*FE* $\pm$ sd, respectively 3.62 ± 1.48 and 0.62 ± 0.73. Further analysis shows that this is not only true for individual instances but for entire groups of constructions. In particular, although their information content is overall higher ($t = 13.511, p < 0.005$, *95% c.i.* $0.371 : 0.497$), referential constructions also have higher facilitating effect than non-referential ones ($t = 3.115, p < 0.005$, *95% c.i.* $0.016 : 0.072$). We conclude that the two measures capture different aspects of a construction's information rate profile, with facilitating effect being sensitive to both construction and utterance information content.

6 The Facilitating Effect of Construction Repetition

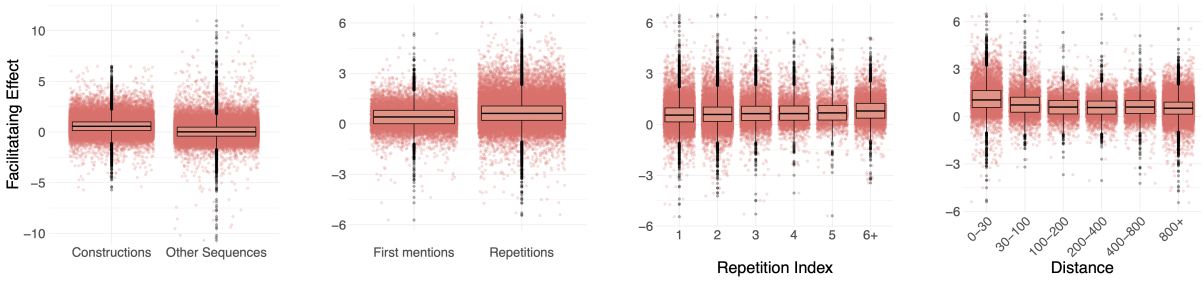
We now test whether constructions have a positive facilitating effect, i.e., whether they reduce the information content of their containing utterances. We present our main statistical model in §6.1, describe the effects of *FE* predictors specific to unique construction mentions in §6.2, and analyse differences between types of constructions in §6.3.

6.1 Method

To understand what shapes a construction's facilitating effect, we collect several of motivated features that can be expected to be informative *FE* predictors. We fit a linear mixed effect (LME) model using (i) these features as fixed effects, (ii) *FE* as the response variable, (iii) and multi-level random effects grouped by dialogue and individual speaker ID. The first predictor is *utterance position*, i.e., the index of the utterance within the dialogue, which allows us to test if *FE* increases over the course of a dialogue. We then include predictors that distinguish different types of repetition. Since we expect a construction mention to increase expectation for subsequent occurrences—thus reshaping their information content—we consider its *repetition index*, i.e., how often the construction has been repeated so far in the dialogue. Expectation is also shaped by intervening material, so we additionally track *distance*, the number of tokens separating a construction mention from the preceding one. As *FE* is the interplay between a construction and its utterance context, it is important to know whether the utterance context contains other mentions of the construction. We use a binary indicator (*previous same utterance*) to single out occurrences whose previous mention is in the same utterance; for these cases, we also count the number of same-utterance previous mentions (*repetition index in utterance*). To explore whether *FE* varies across types of expressions, we also include a binary feature indicating whether the construction is *referential* or non-referential (§3.1). Finally, we keep track of *construction length*, the number of tokens that constitutes a construction,

Speaker	RI	RI Utt	Dist	Turn	$H(u)$	$H(c)$	$FE(c;u)$
A	0	0	-	Drink? that was what he did yeah just just to just to know that I he **might not be** a complete twat but just a fyi	5.99	4.73	0.40
B	1	0	1586	Especially for my birthday mind you I **might not be** here for	5.04	4.01	0.53
	2	1	14	mine and I went what do you mean you **might not be** here?		2.70	0.90

Table 4: Repetition chain for the construction *'might not be'* in dialogue SXWH of the Spoken BNC, annotated with repetition index (RI), RI in utterance (RI Utt), and distance from previous mention (Dist; in tokens). $H(u)$ is the utterance information content, $H(c)$ and $FE(c;u)$ are the construction's information content and facilitating effect.



(a) (b) (c) (d)

Figure 1: The facilitating effect (*FE*) of constructions vs. non-construction sequences (a) and of first construction mentions vs. repetitions (b); as well as *FE* vs. repetition index (c) and *FE* vs. distance from previous mention (number of words). The first distance bin is the mean length of a turn containing a construction (Table 3).

and *PMI*, the pointwise mutual information between a construction and its dialogue, which is essentially a measure of the construction's frequency in the current dialogue as a function of its overall frequency in the corpus, indicating the construction's degree of interaction-specificity.[11]

To determine the fixed effects of the final model, we start with all the predictors listed above (the non-binary ones are log-transformed) and perform backward stepwise selection, iteratively removing the predictor with the lowest significance and keeping only those with $p<0.05$. All predictors make it into our final model, the one which best fits the data according to both the Akaike and the Bayesian Information Criterion. The full specification of the best model, with model fit statistics as well as fixed and random effect coefficients, are in Appendix D. The next two sections present our main findings; we report fixed effect coefficients (β), p-values (p), and 95% confidence intervals (*c.i.*).

6.2 Construction Mentions

Our first observation is that construction usage reduces *utterance* information content. More precisely, we find that **facilitating effect is higher for constructions than for non-construction sequences** ($t = 118.79, p < 0.005$, *95% c.i.* $0.536 : 0.554$). Constructions have on average 62% lower information content than their utterance context; the average percentage drops to 7% for non-construction sequences.[12] Figure 1a shows the two distributions. We also observe a positive effect of utterance position on *FE* ($\beta = 0.046, p < 0.005$, *95% c.i.* $0.026 : 0.06$); that is, **the facilitating effect of constructions increases over the course of dialogues**. While the proportion of construction tokens remains stable (§5.3), their mitigating contribution to utterance information content increases throughout dialogues—perhaps since speakers are more likely to *repeat* established constructions as the dialogue develops. We indeed find that **repeated constructions have stronger facilitating effect**: there is a significant difference between the *FE* of first mentions and repetitions ($t = -38.904, p < 0.005$, *95% c.i.* $-0.265 : -0.239$), as shown in Figure 1b. The information content of repetitions is on average 68% lower than that of their utterance context; for first mentions, it is on average 42% lower.

Having observed that the mitigating contribution of constructions to utterance information content indeed increases with construction repetition, we now look at how the *FE* of repetitions varies as a func-

[11] The probabilities for the PMI calculation are obtained using maximum likelihood estimation over our analysis split of the Spoken BNC.

[12] These are the same sampled non-construction sequences as in §5.2. Their average *FE* is 0.07 ± 0.80.

tion of their distribution across time. On the one hand, we find that **facilitating effect is cumulative**: repeating a construction reduces utterance information content more strongly as more mentions of the construction accumulate in the dialogue (Figure 1c). The effect of repetition index (i.e., how often the construction has been repeated so far in the dialogue) is positive on *FE* ($\beta = 0.079, p < 0.005$, *95% c.i.* $0.063 : 0.094$). On the other hand, the distance of a repetition from the previous mention has a negative effect on *FE* ($\beta = -0.311, p < 0.005$, *95% c.i.* $-0.328 : -0.293$). That is, **facilitating effect decays as a function of the distance between subsequent mentions**. As shown in Figure 1d, this is a fast decay effect: the most substantial drop occurs for low distance values. The large magnitude of this coefficient indicates that recency is an important factor for constructions to have a strong facilitating effect. Indeed, almost one third (31.8%) of all repetitions produced by speakers are not more than 200 tokens apart from their previous mention. Further results showing strong cumulativity effects for self-repetitions within the same utterance can be found in Appendix E.1.

6.3 Types of Construction

In this section, we analyse factors shaping the facilitating effect of construction forms, rather than individual mentions. We focus on the length of a construction and on whether it is referential.

Construction length has a positive effect on *FE* ($\beta = 0.098, p < 0.005$, *95% c.i.* $0.087 : 0.119$): **longer constructions have stronger facilitating effect.** Table 4 shows a full repetition chain for a construction of length 3; Table 5 (Appendix B) for one of length 6. Non-construction sequences display an opposite, weaker trend ($\beta = -0.019, p < 0.05$, *95% c.i.* $-0.032 : -0.005$), as measured with a linear model. A possible explanation for the positive trend of constructions is related to production cost. Longer constructions are more costly for the speaker, so for them to still be an efficient production choice, their facilitating effect must be higher.

Finally, we observe that **referential constructions have a stronger facilitating effect than non-referential ones**. Our LME model yields a positive effect for referentiality on *FE* ($\beta = 0.124, p < 0.005$, 95% c.i $0.099 : 0.149$) and we find a significant difference between the *FE* of the two types ($t = 3.115, p < 0.005$, *95% c.i.* $0.072 : 0.016$). Looking in more detail, first mentions of referential constructions have higher information content and lower *FE* than first mentions of non-referential ones (*H*: $t = 15.435, p < 0.005$, *95% c.i.* $1.115 : 0.864$; *FE*: $t = -9.315, p < 0.005$, *95% c.i.* $-0.246 : -0.161$), perhaps since words in referential sequences tend to be less frequent and more context-dependent. However, when repeated, their information content drops more substantially, reproducing inverse frequency effects attested in humans for syntactic repetitions (Bock, 1986; Scheepers, 2003). As a result, their *FE* exceeds that of non-referential constructions (*FE*: $t = 8.818, p < 0.005$, *95% c.i.* $0.117 : 0.183$), with the information content of a repeated reference being 81% lower than that of its utterance context. Overall, these findings indicate that although referential constructions are less frequent than non-referential ones (23.3% vs. 76.7%; see §3.1), their repetition is a particularly effective strategy of information rate mitigation.

7 Discussion and Conclusions

Construction repetition is a pervasive phenomenon in dialogue; their frequent occurrence gives constructions a processing advantage (Conklin and Schmitt, 2012). In this paper, we show that the processing advantage of constructions can be naturally modelled as reduced information content and propose that speakers' production of constructions can be seen as a strategy for information rate mitigation. This strategy can explain why utterance information content is often found to decrease over the course of dialogues (Vega and Ward, 2009; Giulianelli and Fernández, 2021), in contrast with the predictions of theories of optimal use of the communication channel (Genzel and Charniak, 2002).

We observe that, as predicted, construction usage in English open-domain spoken dialogues mitigates the information rate of utterances. Furthermore, while constructions are produced at a stable rate throughout dialogues, their facilitating effect—our proposed measure of reduction in utterance information content—increases over time. We find that this increment is led by construction repetition, with facilitating effect being positively affected by repetition frequency, density, and by the contents of a construction. Repetitions of referential constructions reduce utterance information content more aggressively, arguably making them a more cost-reducing alternative to the shortening strategy observed in chains of referring expressions (Krauss and

Weinheimer, 1964, 1967), which instead tends to preserve rate constancy (Giulianelli et al., 2021).[13]

Relation to cognitive effort We consider repetitions as a way for speakers to make dialogic interaction less cognitively demanding both on the production and on the comprehension side. This is not at odds with the idea that repetitions are driven by interpersonal synergies (Fusaroli et al., 2014) and coordination (Sinclair and Fernández, 2021). We think that the operationalisation of these higher level processes can be described by means of lower level, efficiency-oriented mechanisms, with synergy and coordination both corresponding to reduced collaborative effort. Although information content estimates from neural language models have been shown to correlate with human processing effort (cf. §2.2), we cannot claim that our work directly models human cognitive processes as we lack the relevant human data to measure such correlation for the corpus at hand.

Adaptive language model Our decision to use an adaptive neural language model affects information content estimates in two main ways. On the one hand, due to their high frequency, constructions are likely to be assigned higher probabilities by this model, and therefore lower information content. We stress that we do not present constructions' lower information content as a novel result, nor do we make any claims based on this result. As explained in §5.2, this is a precondition for our experiments on the facilitating effect of constructions, which is not determined exclusively by their information content (as empirically shown in §5.4) but rather measures the effect of construction usage on the information content of entire utterances. On the other hand, because our model is adaptive, the probability of constructions is likely to increase as a result of their appearance in the dialogue history. Adaptation, however, also contributes to lower utterance information content *overall* through the exploitation of topical and stylistic cues, as demonstrated by the lower perplexity of the adaptive model on the entire target dialogue as well as on other dialogues from the same dataset (see §4.2 and Appendix C.2). In conclusion, while our adaptive language model assigns higher probabilities to frequently repeated tokens—as expected from a psychologically plausible model of utterance processing—it is not responsible for the discovered patterns of construction facilitating effect. In future work, the model can be improved, e.g., by conditioning on the linguistic experience of individual speakers.

Types of dialogue To consolidate our findings, construction repetition patterns should also be studied in dialogues of different genres and on datasets where utterance information content was not found to decrease. We have chosen the Spoken BNC for our study as it contains dialogues from a large variety of real-life contexts, which makes it a representative dataset of open-domain dialogue. In task-oriented dialogue, we expect constructions to consist of a more limited, task-specific vocabulary, resulting in longer chains of repetition and potentially more frequent referential construction usage. These peculiarities of task-oriented dialogue may influence the strength of the facilitating effect (as we have seen, facilitating effect is affected by both frequency and referentiality) but we expect our main results to still hold, as they are generally related to the processing advantage of constructions.

Relevance for dialogue generation models Besides contributing new empirical evidence on construction usage in dialogue, our findings inform the development of more naturalistic utterance generation models. They suggest that models should be continually updated for their probabilities to better reflect human expectations; that attention mechanisms targeting contexts of different sizes (local vs. global) may have a significant impact on the naturalness of generated utterances; and that while anomalous repetitions (e.g., generation loops) should be prevented (Li et al., 2016; Holtzman et al., 2019), it is important to ensure that natural sounding repetitions are not suppressed. We expect dialogue systems that are able to produce human-like patterns of repetitions to be perceived as more natural overall—with users having the feeling that common ground is successfully maintained (Pickering and Garrod, 2004)—and to lead to more effective communication (Reitter and Moore, 2014). In our view, such human-like patterns can be reproduced by steering generation models towards the trends of information rate observed in humans.

Acknowledgements

We would like to thank the members of the Dialogue Modelling Group of the University of Ams-

[13] Expression shortening is more efficient, however, in terms of articulatory cost.

terdam for their useful discussions, and our anonymous reviewers for their insightful comments. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455).

References

Inbal Arnon and Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of memory and language*, 62(1):67–82.

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):31–56.

Matthew Aylett and Alice Turk. 2006. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, 119(5):3048–3058.

Colin Bannard and Danielle Matthews. 2008. Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological science*, 19(3):241–248.

Alan Bell, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory, and Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, 113(2):1001–1024.

Lasse Bergroth, Harri Hakonen, and Timo Raita. 2000. A survey of longest common subsequence algorithms. In *Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000*, pages 39–48. IEEE.

Douglas Biber and Federica Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for specific purposes*, 26(3):263–286.

Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. If you look at...: Lexical bundles in university teaching and textbooks. *Applied linguistics*, 25(3):371–405.

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):1347–1375.

J Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–1493.

Joan Bybee. 2006. From usage to grammar: The mind's response to repetition. *Language*, pages 711–733.

Joan Bybee. 2010. *Language, usage and cognition*. Cambridge University Press.

Joan Bybee and Joanne Scheibman. 1999. The effect of usage on degrees of constituency: The reduction of don't in English. *Linguistics*, 37(4):575–596.

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):95–122.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

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

Meghan Clayards, Michael K. Tanenhaus, Richard N. Aslin, and Robert A. Jacobs. 2008. Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3):804–809.

Georgie Columbus. 2013. In support of multiword unit classifications: Corpus and human rating data validate phraseological classifications of three different multiword unit types. *Yearbook of Phraseology*, 4(1):23–44.

Kathy Conklin and Norbert Schmitt. 2012. The processing of formulaic language. *Annual Review of Applied Linguistics*, 32:45–61.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Vera Demberg, Asad Sayeed, Philip Gorinski, and Nikolaos Engonopoulos. 2012. Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 356–367.

Nina Dethlefs, Helen Hastie, Heriberto Cuayáhuitl, Yanchao Yu, Verena Rieser, and Oliver Lemon. 2016. Information density and overlap in spoken dialogue. *Computer speech & language*, 37:82–97.

Gabriel Doyle and Michael Frank. 2015a. Shared common ground influences information density in microblog texts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1587–1596, Denver, Colorado. Association for Computational Linguistics.

Gabriel Doyle and Michael C. Frank. 2015b. Audience size and contextual effects on information density in Twitter conversations. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 19–28.

Guillaume Dubuisson Duplessis, Franck Charras, Vincent Letard, Anne-Laure Ligozat, and Sophie Rosset. 2017a. Utterance retrieval based on recurrent surface text patterns. In *European Conference on Information Retrieval*, pages 199–211. Springer.

Guillaume Dubuisson Duplessis, Chloé Clavel, and Frédéric Landragin. 2017b. Automatic measures to characterise verbal alignment in human-agent interaction. In *18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 71–81.

Guillaume Dubuisson Duplessis, Caroline Langlet, Chloé Clavel, and Frédéric Landragin. 2021. Towards alignment strategies in human-agent interactions based on measures of lexical repetitions. *Language Resources and Evaluation*, 55(2):353–388.

Austin F. Frank and T. Florian Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Riccardo Fusaroli, Joanna Rączaszek-Leonardi, and Kristian Tylén. 2014. Dialog as interpersonal synergy. *New Ideas in Psychology*, 32:147–157.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 199–206.

Dmitriy Genzel and Eugene Charniak. 2003. Variation of entropy and parse trees of sentences as a function of the sentence number. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 65–72.

Edward Gibson, Leon Bergen, and Steven T. Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056.

Mario Giulianelli and Raquel Fernández. 2021. Analysing human strategies of information transmission as a function of discourse context. In *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2021. Is information density uniform in task-oriented dialogues? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Adele E Goldberg. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

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 (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Daniel S Hirschberg. 1977. Algorithms for the longest common subsequence problem. *Journal of the ACM (JACM)*, 24(4):664–675.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

T. Florian Jaeger and Roger P. Levy. 2007. Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems*, pages 849–856.

Niels Janssen and Horacio A Barber. 2012. Phrase frequency effects in language production. *PloS one*, 7(3):e33202.

Frederick Jelinek, Lalit Bahl, and Robert Mercer. 1975. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 21(3):250–256.

Hajnal Jolsvai, Stewart M McCauley, and Morten H Christiansen. 2013. Meaning overrides frequency in idiomatic and compositional multiword chunks. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Frank Keller. 2004. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 317–324.

Ben Krause, Emmanuel Kahembwe, Iain Murray, and Steve Renals. 2018. Dynamic evaluation of neural sequence models. In *International Conference on Machine Learning*, pages 2766–2775. PMLR.

Robert M Krauss and Sidney Weinheimer. 1964. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1(1):113–114.

Robert M Krauss and Sidney Weinheimer. 1967. Effect of referent similarity and communication mode on verbal encoding. *Journal of Verbal Learning and Verbal Behavior*, 6(3):359–363.

Koenraad Kuiper. 1995. *Smooth talkers: The linguistic performance of auctioneers and sportscasters*. Routledge.

Roger Levy. 2008. A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 234–243.

Roger Levy, Klinton Bicknell, Tim Slattery, and Keith Rayner. 2009. Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106(50):21086–21090.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Robbie Love, Claire Dembry, Andrew Hardie, Vaclav Brezina, and Tony McEnery. 2017. The spoken BNC2014. *International Journal of Corpus Linguistics*, 22(3):319–344.

Danny Merkx and Stefan L Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22.

Irene Fernandez Monsalve, Stefan L Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408.

M. J. Pickering and S. Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02):169–190.

Ting Qian and T. Florian Jaeger. 2011. Topic shift in efficient discourse production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

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

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*, pages 121–124.

David Reitter and Johanna D Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76:29–46.

Christoph Scheepers. 2003. Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition*, 89(3):179–205.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45).

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Arabella Sinclair and Raquel Fernández. 2021. Construction coordination in first and second language acquisition. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Potsdam, Germany. SEMDIAL.

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*, 175:111–122.

Anna Siyanova-Chanturia, Kathy Conklin, and Walter JB Van Heuven. 2011. Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(3):776.

Patrizia Tabossi, Rachele Fanari, and Kinou Wolf. 2009. Why are idioms recognized fast? *Memory & Cognition*, 37(4):529–540.

Debra Titone and Maya Libben. 2014. Time-dependent effects of decomposability, familiarity and literal plausibility on idiom priming: A cross-modal priming investigation. *The Mental Lexicon*, 9(3):473–496.

Debra A Titone and Cynthia M Connine. 1994. Descriptive norms for 171 idiomatic expressions: Familiarity, compositionality, predictability, and literality. *Metaphor and Symbol*, 9(4):247–270.

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

Antoine Tremblay and R Harald Baayen. 2010. Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. *Perspectives on formulaic language: Acquisition and communication*, pages 151–173.

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.

Geoffrey Underwood, Norbert Schmitt, and Adam Galpin. 2004. The eyes have it: An eye movement study into the processing of formulaic sequences. In Norbert Schmitt, editor, *Formulaic Sequences: Acquisition, Processing and Use*, pages 153–172. John Benjamins.

Marten van Schijndel and Tal Linzen. 2018. A neural model of adaptation in reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4704–4710.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30:5998–6008.

Alejandro Vega and Nigel Ward. 2009. Looking for entropy rate constancy in spoken dialog. Technical Report UTEP-CS-09-19, University of Texas El Paso.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Alison Wray. 2002. *Formulaic Language and the Lexicon*. Cambridge, UK: Cambridge University Press.

Yadong Xi, Jiashu Pu, and Xiaoxi Mao. 2021. Taming repetition in dialogue generation. *arXiv preprint arXiv:2112.08657*.

Yang Xu and David Reitter. 2018. Information density converges in dialogue: Towards an information-theoretic model. *Cognition*, 170:147–163.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DialoGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Appendix

A Possible Criteria to Distinguish Constructions

Lexicalised constructions can be classified according to multiple criteria (Titone and Connine, 1994; Wray, 2002; Columbus, 2013), including those listed below.

- **Compositionality** This criterion is typically used to separate idioms from other formulaic expressions, although it is sometimes referred to as *transparency* to underline its graded, rather than binary, nature. There is no evidence, however, that the processing advantage of idioms differs from that of compositional phrases (Tabossi et al., 2009; Jolsvai et al., 2013; Carrol and Conklin, 2020). *Therefore we ignore this criterion in the current study.*

- **Literal plausibility** This criterion is typically used to discriminate among different types of idioms (Titone and Connine, 1994; Titone and Libben, 2014)—as compositional phrases are literally plausible by definition. *Because we ignore distinctions made on the basis of compositionality, we do not use this criterion.*

- **Meaningfulness** Meaningful expressions are idioms and compositional phrases (e.g. *'on my mind'*, *'had a dream'*) whereas sentence fragments that break constituency boundaries (e.g., *'of a heavy'*, *'by the postal'*) are considered less meaningful (as measured in norming studies, e.g., by Jolsvai et al., 2013). There is some evidence that the meaningfulness of multi-word expressions correlates with their processing advantage even more than their frequency (Jolsvai et al., 2013); yet expressions are particularly frequent, they present processing advantages even if they break regular phrasal structures (Bybee and Scheibman, 1999; Tremblay et al., 2011). Moreover, utterances that break regular constituency rules are particularly frequent in spoken dialogue data (e.g., *'if you could search for job and that's not'*, *'you don't wanna damage your relationship with'*). *For these reasons, we do not exclude constructions that span multiple constituents from our analysis.*

- **Schematicity** This criterion distinguishes expressions where all the lexical elements are fixed from expressions "with slots" that can be filled by varying lexical elements.*In this study, we focus on fully lexicalised constructions.*

- **Familiarity** This is a subjective criterion that strongly correlates with objective frequency

(Carrol and Conklin, 2020). Human experiments would be required to obtain familiarity norms for our target data, and the resulting norms would only be an approximation of the familiarity judgements of the true speakers we analyse the language of. *Therefore, we ignore this criterion in the current study.*

- **Communicative function** Formulaic expressions can fulfil a variety of discourse and communicative functions. Biber et al. (2004), e.g., distinguish between stance expressions (attitude, certainty with respect to a proposition), discourse organisers (connecting prior and forthcoming discourse), and referential expressions; and for each of these three primary discourse functions, more specific subcategories are defined. This type of classification is typically done a posteriori—i.e., after a manual analysis of the expressions retrieved from a corpus according to other criteria (Biber and Barbieri, 2007). In the BNC, for example, we find epistemic lexical bundles (*'I don't know'*, *'I don't think'*), desire bundles (*'do you want to'*, *'I don't want to'*), obligation/directive bundles (*'you don't have to'*), and intention/prediction bundles (*'I'm going to'*, *'it's gonna be'*). *We do not use this criterion to avoid an a priori selection of the constructions.*

B Extraction of Repeated Constructions

We define a limited specific vocabulary of generic nouns that should not be considered referential. The vocabulary includes: *bit*, *bunch*, *day*, *days*, *fact*, *god*, *idea*, *ideas*, *kind*, *kinds*, *loads*, *lot*, *lots*, *middle*, *ones*, *part*, *problem*, *problems*, *reason*, *reasons*, *rest*, *side*, *sort*, *sorts*, *stuff*, *thanks*, *thing*, *things*, *time*, *times*, *way*, *ways*, *week*, *weeks*, *year*, *years*. We also find all the filled pauses and exclude word sequences that consist for more than 50% of filled pauses. Filled pauses in the Spoken BNC are transcribed as: *huh*, *uh*, *erm*, *hm*, *mm*, *er*.

Figure 2 shows the proportion of tokens in an utterance belonging to constructions (referential and non-referential) and to non-construction sequences. Table 5 shows a whole construction chain (from the first mention to the last repetition) for a construction of length 6.

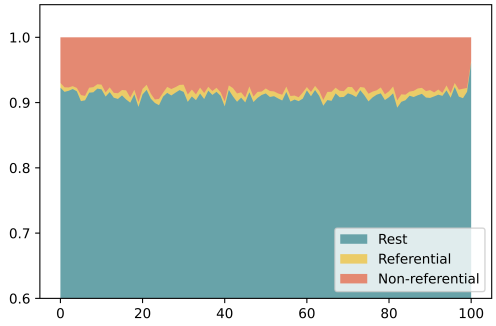


Figure 2: Proportion of tokens in an utterance that belong to referential constructions, non-referential constructions, and to non-construction sequences. The x axis shows percentages indicating utterance positions in the dialogue relative to the dialogue length.

C Language Model

C.1 Finetuning

We finetune the *'small' variant* of GPT-2 (Radford et al., 2019) and DialoGPT (Zhang et al., 2020) on our finetuning split of the Spoken BNC (see Section 3) using HuggingFace's implementation of the models with default tokenizers and parameters (Wolf et al., 2020). Dialogue turns are simply concatenated; we have experimented with labelling the dialogue turns (i.e., *A: utterance 1, B: utterance 2* and found that this leads to higher perplexity. The finetuning results for both models are presented in Table 6. We finetune the models and measure their perplexity using Huggingface's finetuning script. We use early stopping over 5 epochs.[14] Sequence length and batch size vary together because they together determine the amount of memory required; more expensive combinations (e.g., 256 tokens with batch size 16) require an exceedingly high amount of GPU memory. Reducing the maximum sequence length has limited impact: 99.90% of dialogue turns have at most 128 words.

DialoGPT starts from extremely high perplexity values but catches up quickly with finetuning. GPT-2 starts from much lower perplexity values and reaches virtually the same perplexity as DialoGPT after finetuning. For the pre-trained DialoGPT per-

[14]The number of epochs (5) has been selected in preliminary experiments together with the learning rate ($1e-4$). In these experiments—which we ran for 40 epochs—we noticed that the $1e-4$ learning rate offers the best tradeoff of training time and perplexity out of four possible values: $1e{-}2$, $1e{-}3$, $1e{-}4$, $1e-5$. We obtained insignificantly lower perplexity values with a learning rate of $1e-5$, with significantly longer training time: 20 epochs for GPT-2 and 28 epochs for DialoGPT.

Speaker	RI	RI Utt	Dist	Turn	$H(u)$	$H(c)$	$FE(c,u)$
A	0	0	-	[...] I think that everyone should have the same opportunities and **I don't think you should be** proud or ashamed of what your you know what your situation is whether you what your what your race is whether you're a woman or a man whether you live from this pl whether you're in this place [...]	4.24	1.90	1.21
A	1	0	80	I well I th I don't think it should **I don't think you should be**	3.40	1.73	1.40
A	2	0	19	Well yes perhaps but **I don't think you should be** like um embarrassed about it or I think I think you should just sort of	3.95	1.06	2.25

Table 5: Repetition chain for the construction *'I don't think you should be'* in dialogue S2AX of the Spoken BNC, annotated with repetition index (RI), repetition index in utterance (RI Utt), and distance from previous mention (Dist; number of tokens). $H(u)$ is the utterance information content, $H(c)$ and $FE(c,u)$ are the construction's information content and facilitating effect.

plexity is extremely high, and the perplexity trend against maximum sequence length is surprisingly upward. These two behaviours indicate that the pretrained DialoGPT is less accustomed than GPT-2 to the characteristics of our dialogue data. DialoGPT is trained on written online group conversations, while we use a corpus of transcribed spoken conversations between two speakers. In contrast, GPT-2 has been exposed to the genre of fiction, which contains scripted dialogues, and thus to a sufficiently similar language use. We select GPT-2 finetuned with a maximum sequence length of 128 and 512 as our best two models; these two models (which we now refer to as *frozen*) are used for the adaptive learning rate selection (Section C.2).

C.2 Learning Rate Selection

To find the appropriate learning rate for on-the-fly adaptation (see Section 4.2), we randomly select 18 dialogues D from the analysis split of the Spoken BNC and run an 18-fold cross-validation for a set of six candidate learning rates: $1\mathrm{e}-5$, $1\mathrm{e}-4$, ..., 1. We finetune the model on each dialogue using one of these learning rate values, and compute perplexity change 1) on the dialogue itself (to measure *adaptation*) as well as 2) on the remaining 17 dialogues (to measure *generalisation*). We set the Transformer's context window to 50 to reproduce the experimental conditions presented in Section 4.1.

More precisely, for each dialogue $d \in D$, we calculate the perplexity of our two frozen models (Section C.1) on d and $D \setminus \{d\}$ (which we refer to as $ppl_{before}(d)$ and $ppl_{before}(D)$, respectively). Then, we finetune the models on d using the six candidate learning rates, and measure again the perplexity over d and $D \setminus \{d\}$ (respectively, $ppl_{after}(d)$ and $ppl_{after}(D)$). The change in performance is evaluated according to two metrics: $\frac{ppl_{after}(d)-ppl_{before}(d)}{ppl_{before}(d)}$ measures the degree to which the model has successfully adapted to the target dialogue; $\frac{ppl_{after}(D)-ppl_{before}(D)}{ppl_{before}(D)}$ measures whether finetuning on the target dialogue has caused any loss of generalisation.

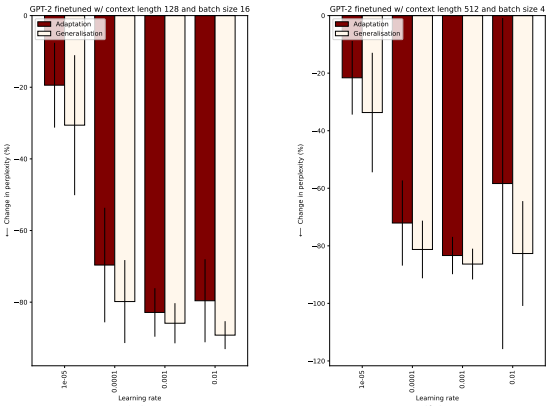
The learning rate selection results are presented in Figure 3. We select $1\mathrm{e}-3$ as the best learning rate and pick the model finetuned with a maximum sequence length of 512 as our best model. The difference in perplexity reduction (both adaptation and generalisation) is minimal with respect to the model finetuned with a maximum sequence length of 128, but since the analysis split of the Spoken BNC contains turns longer than 128 tokens, we select the 512 version. Similarly to van Schijndel and Linzen (2018), we find that finetuning on a dialogue does not cause a loss in generalisation but instead helps the model generalise to other dialogues. Unlike (2018), who used LSTM language models, we find that learning rates larger than $1\mathrm{e}-1$ cause backpropagation to overshoot, even within a single dialogue. In Figure 3, the bars for $1\mathrm{e}-1$ and 1 are not plotted because the corresponding data contains infinite perplexity values (due to numerical overflow). The selected learning rate, $1\mathrm{e}-3$, is a relatively low learning rate for on-the-fly adaptation but it is still higher than the best learning rate for the entire dataset by a factor of 10.

D Linear Mixed Effect Models

As explained in §6.1 of the main paper, we fit a linear mixed effect model using facilitating effect as the response variable and including multilevel random effects grouped by dialogues and individ-

Model	Learning rate	Max sequence length	Batch size	Best epoch	Perplexity finetuned	Perplexity pre-trained
DialoGPT	0.0001	128	16	3	23.21	7091.38
DialoGPT	0.0001	256	8	4	22.26	12886.92
DialoGPT	0.0001	512	4	4	21.73	21408.32
GPT-2	0.0001	128	16	4	23.32	173.76
GPT-2	0.0001	256	8	3	22.21	159.23
GPT-2	0.0001	512	4	3	21.55	149.82

Table 6: Finetuning results for GPT-2 and DialoGPT on our finetuning split of the Spoken BNC.



(a) (b)

Figure 3: The adaptation and generalisation performance (defined in Section C.2) with varying learning rate.

ual speakers.[15]. The fixed effects of the model, resulting from a backward stepwise selection procedure, are presented in §6.1. Non-binary predictors are log-transformed, mean-centered, and scaled by 2 sd. The final model is summarised in Listing 1 and its coefficients are visualised in Figure 4. We rely on the `lme4` and `lmerTest` R packages for this analysis.

E Further Results

E.1 Same-Utterance Self-Repetitions

We investigate the interaction between cumulativity and recency (see §6.2) by focusing on densely clustered repetitions, produced by a speaker within a single utterance (the median distance between repetitions in the same utterance is 8 words; across turns it is 370.5 words). Table 4 shows an example of same-utterance repetition. Repeating a construction when it has already been mentioned in the current utterance limits its facilitating effect ($\beta = -0.099, p < 0.05$, *95% c.i.* -0.184:-0.013): if a portion of the utterance already consists of a construction, utterance information content will already be reduced, which in turn reduces the potential for the facilitating effect of repetitions. Nevertheless, we find **strong cumulativity effects for self-repetitions within the same utterance**: the repetition index *within the current utterance* of a construction mention (i.e., how often the construction has been repeated so far in the utterance) has a positive effect on *FE* ($\beta = 0.178, p < 0.005$, *95% c.i.* 0.130:0.226); see Figure 5a. In sum, same-utterance self-repetitions, especially those involving three or more mentions in a single utterance, can have a strong reduction effect on utterance information content. Although this may seem a simple yet very effective strategy for information rate mitigation, it is unlikely to be very effective in terms of the amount of information exchanged. Indeed, speakers do not use this strategy often in the Spoken BNC: 6.82% of the total construction occurrences have at least one previous mention in the same utterance.

E.2 Interaction-Specificity

To distinguish interaction-specific constructions—those repeated particularly often in certain dialogues—from interaction-agnostic ones, we measure the association strength between a construction c and a dialogue d as the pointwise mutual information (PMI) between the two:

$$\mathrm{PMI}(c, d) = \log_2 \frac{P(c|d)}{P(c)} \quad [5]$$

This quantifies how unusually frequent a construction is in a given dialogue, compared to the rest of the corpus. For example, for a construction to obtain a PMI score of 1, its probability given the dialogue $P(c|d)$ must be twice as high as its prior probability $P(c)$. Low PMI scores (especially below 1) characterise interaction-agnostic constructions, whereas higher PMI scores indicate

[15]We also try grouping observations only by dialogue and only by individual speakers. The amount of variance explained (but unaccounted for by the fixed effects) decreases, so we keep the two-level random effects.

Listing 1: Linear mixed effect model for Facilitating Effect

```
MODEL INFO:
Observations: 46399
Dependent Variable: Facilitating Effect
Type: Mixed effects linear regression

MODEL FIT:
AIC = 99197.283, BIC = 99302.224
Pseudo-R^2 (fixed effects) = 0.084
Pseudo-R^2 (total) = 0.111

FIXED EFFECTS:
```

	Est.	2.5%	97.5%	t val.	d.f.	p
(Intercept)	0.704	0.683	0.725	65.527	185.698	0.000
log Utterance Position	0.046	0.026	0.066	4.556	9274.269	0.000
log Construction Length	0.098	0.084	0.111	14.396	46372.022	0.000
log Repetition Index	0.079	0.063	0.094	10.096	45082.205	0.000
log Distance	-0.311	-0.328	-0.293	-34.571	46269.156	0.000
Previous Same Utterance	-0.099	-0.184	-0.013	-2.262	46063.723	0.024
log Rep. Index in Utterance	0.178	0.130	0.226	7.243	45765.367	0.000
PMI	-0.139	-0.154	-0.124	-18.225	45172.205	0.000
Referential	0.124	0.099	0.149	9.887	46214.616	0.000

```
p values calculated using Satterthwaite d.f.

RANDOM EFFECTS:
```

Group	Parameter	Std. Dev.
Speaker:‘Dialogue ID	(Intercept)	0.082
Dialogue ID	(Intercept)	0.090
Residual		0.701

```
Grouping variables:
```

Group	# groups	ICC
Speaker:‘Dialogue ID	368	0.013
Dialogue ID	185	0.016

```
Continuous predictors are mean-centered and scaled by 2 s.d.
```

that constructions are specific to a given dialogue. The probabilities in Eq. 5 are obtained using maximum likelihood estimation over the analysis split of the Spoken BNC. PMI scores have a negative effect on *FE* ($\beta = -0.139, p < 0.005$, 95% c.i. -0.154:-0.124), indicating that interaction-agnostic constructions have a stronger facilitating effect than interaction-specific ones. Figure 5b shows the *FE* distributions for the most extreme cases: constructions with a PMI lower than 1 ('agnostic') and constructions that have been repeated in only one dialogue ('specific').

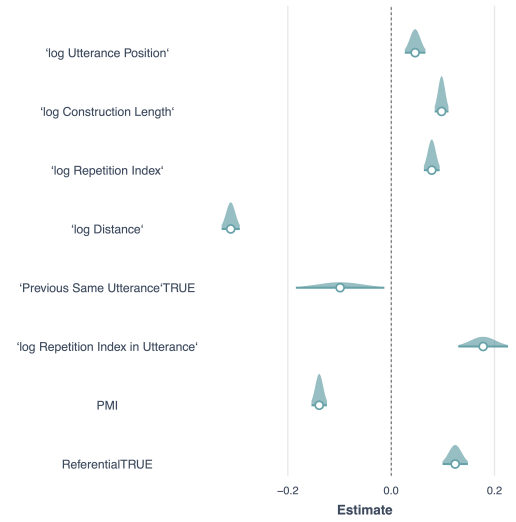
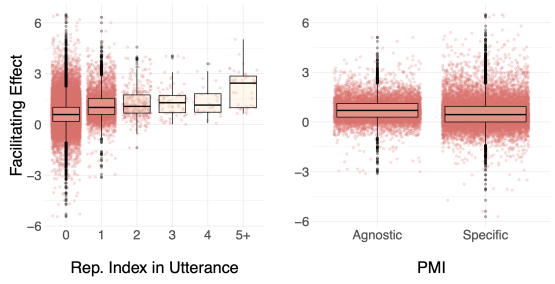


Figure 4: Significant predictors of facilitating effect. Mixed effects linear regression, continuous predictors are mean-centred and scaled by 2 standard deviations.



(a) (b)

Figure 5: Facilitating effect against repetition index *within the current utterance* (a) and facilitating effect of interaction-agnostic constructions ($\mathrm{PMI}(c, d) < 1$) vs. interaction-specific constructions ($\mathrm{PMI}(c, d) = \max_{c',d'} \mathrm{PMI}(c', d')$) (b).

F Computing Infrastructure and Budget

Our experiments were carried out using a single GPU on a computer cluster with Debian Linux OS. The GPU nodes on the cluster are GPU GeForce 1001 1080Ti, 11GB GDDR5X, with NVIDIA driver version 418.56 and CUDA version 10.1. The total computational budget required to finetune the language model amounts to 45 minutes; obtaining surprisal estimates requires 4 hours, and selecting the adaptation learning rate requires 9 hours.