Repetition Facilitates Processing: The Processing Advantage of Construction Repetition in Dialogue

Anonymous ACL submission

Abstract

Repetitions occur frequently in dialogue. This study focuses on the repetition of lexicalised constructions—i.e., recurring multi-word units—in English open domain spoken dialogues. We hypothesise that construction repetition is an efficient communication strategy that reduces processing effort, and we make three predictions based on this hypothesis. We conduct a quantitative analysis, measuring reduction in processing effort via two surprisal-based measures and estimating surprisal with an adaptive neural language model. Our three predictions are confirmed: (i) repetitions facilitate the processing of constructions and of their linguistic context; (ii) facilitating effects are higher when repetitions accumulate, (iii) and lower when repetitions are less locally distributed. Our findings suggest that human-like patterns of repetitions can be learned implicitly by utterance generation models equipped with psycholinguistically motivated learning objectives and adaptation mechanisms.

1 Introduction

In language production, speakers select—among a set of possible realisations—the lexical, syntactic, and semantic alternatives they deem most appropriate to verbalise their communicative intents. For instance, speakers can choose to precede reported speech with *'I said'* or *'I was like'*: *'I was like where is this going?'*, *'I said you don't have to love each other'*. Given such sets of alternatives, speakers' choices are influenced, among other things, by their recent linguistic experience. In a dialogue, a speaker may be more prone to choose *'I was like'* if they or their conversational partner have already used it. This is an example of priming: under the influence of previous mentions, *'I was like'* is repeated more often than expected by chance.

Most studies on priming have targeted the repetition of syntactic structures (Levelt and Kelter, 1982; Bock, 1986; Branigan et al., 2000; Reitter et al., 2006b, 2011), often explaining them within the framework of the interactive alignment model (Pickering and Garrod, 2004). Lexical repetitions have also been investigated (e.g., Brennan, 1996) and they have been typically explained as the result of collaborative mechanisms (Brennan and Clark, 1996) or social pressures (Danescu-Niculescu-Mizil et al., 2012; Noble and Fernández, 2015; Doyle and Frank, 2016). Less is known about the mechanisms underlying speakers' repetition of particular configurations of structures and lexemes, *constructions*, a pervasive phenomenon in conversational language use (Tomasello, 2003; Goldberg, 2006; Sinclair and Fernández, 2021). The reuse of constructions has been analysed by Fusaroli et al. (2014) as part of a process of 'interpersonal synergy' between conversational partners. In this study, we investigate whether speakers repeat lexicalised constructions (such as *'I was like'*) throughout a dialogue as a result of two information processing mechanisms traditionally argued to affect priming: 1) *residual activations* due to exposure to local context (Pickering and Branigan, 1998; Cleland and Pickering, 2003) and 2) *implicit learning* of the global statistics of expressions and structures (Bock and Griffin, 2000; Fine and Florian Jaeger, 2013). We use a computational model to approximate these mechanisms, hypothesising that, if they are in place, construction repetition becomes a rational strategy of information transmission (Gibson, 1998; Levy, 2008): processing effort is reduced when speakers follow this strategy.

We use *surprisal* to operationalise the processing advantage of construction repetition, estimated with a neural language model. Surprisal measures the unpredictability of a linguistic signal, which can be taken as an estimate of the amount of effort required to process the signal (e.g., Jelinek et al., 1975; Keller, 2004; Levy, 2008). We predict (i) that construction repetition has a facilitating effect on processing, observable in the form of a surprisal reduction both for the construction itself and for its linguistic context. To further understand the nature of the processing advantage, we study how it varies across different types of repetition. We predict (ii) that the processing advantage of construction repetition increases with the total number of repetitions made in a dialogue, and (iii) that it decreases with the distance between repetitions. Our experiments confirm these three predictions, providing new empirical evidence that dialogue partners use repetitions as a communication strategy due to it leading to higher information processing efficiency.

Our findings inform the development of better dialogue models. They indicate that avoiding repetitions in utterance generation (Li et al., 2016; Welleck et al., 2019) may not be the most appropriate strategy. Instead, models should be encouraged to follow human-like patterns of repetitions to be successfully deployed in conversational settings.

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 Section 5, 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, 2004; Siyanova-Chanturia et al., 2011), and electrophysiology (Tremblay and Baayen, 2010; Siyanova-Chanturia et al., 2017). In this paper, we study the processing advantage of the *repetition* of lexicalised constructions.

SYXU	S7ZG	SVPK
had a few	*if you look at*	*I think it was just*
it I was	*yes of course*	*like this is*
I'd be like	*look at what*	*like you're not*
were like oh	*if you give*	*so I didn't*
do you get	*and all of that*	*that I know*
and I went	*it doesn't have to*	*it's not even*
I don't like	*right okay so*	*and I was kind of*
a bit more	*something out of*	*and it was like oh*
I know I	*that in itself*	*think of it like*
I was like	*yeah that's fine*	*kind of thing where*

Table 1: Top 10 constructions from three dialogues of the Spoken BNC (Love et al., 2017). Constructions are sorted according to the PMI between a construction and its dialogue (see Section 5 for extraction procedure). Headers correspond to the dialogues' IDs in the corpus.

2.2 Surprisal and Processing Effort

Estimates of surprisal 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). Because 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 an optimal strategy to manage the fluctuations of surprisal levels 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 uniform surprisal rate in texts (Genzel and Charniak, 2002, 2003; Qian and Jaeger, 2011).

To estimate surprisal, we use GPT-2 (Radford et al., 2019), a neural language model. Using language models to approximate surprisal is an established approach (e.g., Genzel and Charniak, 2002; Keller, 2004; Xu and Reitter, 2018) and *neural* models' surprisal estimates in particular have been shown to be good predictors of processing effort, measured as reading time, gaze duration, and N400 response (van Schijndel and Linzen, 2018; Merkx and Frank, 2021).

2.3 Priming Mechanisms

Priming has been widely studied through the analysis of structural repetitions, whether densely clustered (e.g., Branigan et al., 1999; Wheeldon and Smith, 2003), or occurring across multiple utterances and interactions (e.g., Branigan et al., 2000; Kaschak et al., 2014). These two types of priming (often called *short-term priming* and *long-term priming*, respectively) are thought to be the result of different underlying mechanisms (for a review see, e.g., Hartsuiker et al., 2008). Quickly decaying, short-term priming effects rely on an activation-based mechanism dependent on residual traces left by lexical material (Pickering and Branigan, 1998; Cleland and Pickering, 2003). Slowly decaying, long-term priming effects are independent of lexical material and rely on an implicit learning mechanism (Bock and Griffin, 2000; Fine and Florian Jaeger, 2013). In the current study, we model both mechanisms so that we do not limit a priori the space of possible processes underlying priming.

3 Hypotheses

Does construction repetition come with a processing advantage? Is this advantage due to the mechanisms underlying priming? To answer these questions, we formulate the following three hypotheses.

H1 *Repetition facilitates processing.* We predict 1) a construction has lower surprisal when repeated than when first produced, and 2) repetitions of a construction (i.e., the occurrences that follow its first mention) have a stronger reduction effect on the surprisal of the dialogue turn (i.e., a stronger *facilitating effect*) than first mentions.

H2 *The processing advantage of repetition is cumulative.* We predict multiple repetitions of a construction contribute 1) to a stronger reduction in the surprisal of the construction itself, and 2) to a stronger facilitating effect.

H3 *The processing advantage of repetition decays as a function of the distance between repetitions*. We predict that a larger distance between a construction repetition and its previous mention results 1) in a weaker reduction in the surprisal of the construction, and 2) in a weaker facilitating effect.

H1 tests whether repeating a construction reduces processing effort. Comprehenders are known to process written and spoken words more rapidly when they are repeated (for a review, see Bigand et al., 2005), suggesting increased expectation for these words. An increase in expectation (hence reduction in surprisal) due to repetition is compatible with the implicit learning account of priming (Kaschak et al., 2006; Reitter et al., 2011; Fine et al., 2013). However, if repetitions are closely clustered, any surprisal reduction could also be the result of residual activations from previous mentions, in line with the activation-based account.

Because **H1** does not distinguish between different repetitions of a construction and their distribution across time, **H2** tests how surprisal reduction effects vary along chains of repetitions in terms of cumulation (Table 4 shows an example chain). Changes in the magnitude of the processing advantage of construction repetition may interact with the number of times the construction has already been repeated (Jaeger and Snider, 2008; Fine and Jaeger, 2016). Cumulative effects propagating over distant repetitions would be evidence in favour of the implicit learning account, whereas cumulative effects taking place locally are compatible with the activation-based account.

The processing advantage of construction repetition may also be determined by the distance between mentions. Inspired by earlier analyses conducted for lexical and syntactic priming with varying results (Reitter et al., 2011; Howes et al., 2010; Healey et al., 2014), **H3** investigates the influence of recency of previous mention on a repetition's processing advantage. Fast decay effects could be taken in support of the activation-based account, whereas slow decay effects would suggest reduction in surprisal is due to sensitivity to the global statistics of expressions and structures in a dialogue, in line with the implicit learning account.

4 Data

We test our hypotheses on the Spoken British National Corpus[1] (Love et al., 2017), a dataset of transcribed spoken open domain 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 Section 6) and a 30%

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

analysis set. Table 2 shows basic statistics for the dialogues used in this study.

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

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

5 Extracting Repeated Constructions

We define constructions as multi-word sequences that are repeated within a dialogue. We analyse constructions produced by only one of the dialogue participants as well as those produced by both speakers. To extract a set of 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 dialogue turn. We focus on constructions of at least three tokens, uttered at least three times in a dialogue. Repeated sequences that mostly appear as a sub-part of a larger repeated construction are discarded.[3]

We apply the following further constraints. First, we exclude topic-determined constructions and referential expressions in order to disentangle priming effects from topic coherence effects. To this end, we filter out constructions that include nouns, unless the nouns are highly generic.[4] For example, we discard sequences such as *'playing table tennis'* and *'a woolly jumper'* and retain constructions such as *'a lot of'* and *'the thing is'*. Second, we filter out repetitions that are simply due to a high base frequency rate and not to the speakers' self and mutual priming effects. We measure the association strength between a construction c and a dialogue d as the pointwise mutual information (PMI) between the two:

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

which measures how unusually frequent a construction is in a given dialogue, compared to the rest of the corpus. We discard all constructions that have a PMI score lower than 1 in their respective dialogue. The probabilities in Eq. 1 are obtained using maximum likelihood estimation over the analysis split of the Spoken BNC. Finally, we exclude sequences containing punctuation marks or which consist of more than 50% filled pauses (e.g., *'mm'*, *'erm'*).[5]

Applying the described extraction procedure to the 187 dialogues in the analysis split of the Spoken BNC, we obtain a total of 3,676 unique constructions and 33,103 occurrences. Further statistics on the extracted constructions are presented in Table 3. Table 1 shows examples of the top 10 constructions extracted from three dialogues, ranked according to their PMI score.

	Mean ± Std	Median	Min	Max
Construction length	3.23 ± 0.52	3	3	7
Construction frequency	3.87 ± 1.93	3	3	58
Constructions per dialogue	206 ± 307	100	3	2023
Words per dialogue turn	31 ± 37	21	3	959

Table 3: Construction statistics for the analysis split of the Spoken BNC. *Construction frequency* is the number of occurrences of a given construction in a dialogue, *Constructions per dialogue* is the number of occurrences of all constructions in a dialogue, *Words per dialogue turn* is computed on turns containing a construction.

6 Experimental Setup

In this section, we present two surprisal-based measures of processing advantage, the adaptive language model that produces surprisal estimates, and statistical tests used to confirm our hypotheses.[6]

6.1 Measures of Processing Advantage

The *surprisal* of a word choice w_i is the negative logarithm of the corresponding word probability, conditioned on the dialogue turn context t (i.e., the words that precede w_i in the dialogue turn) and on the local dialogue context l:

$$H(w_i|t,l) = -\log_2 P(w_i|t,l) \qquad [2]$$

We define the local dialogue context l as the 50 tokens that precede the first word in the dialogue turn.[7] We use tokens as a unit of context size, rather

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

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

[6]Data and code will be made public upon acceptance.

[7]Building on prior work (Reitter et al., 2006a) that uses a window of 15 seconds of spoken dialogue as the locus of

Speaker	RI	RI Turn	Dist	Turn	S	FE
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	4.73	0.40
B	1 2	0 1	1586 14	Especially for my birthday mind you I **might not be** here for mine and I went what do you mean you **might not be** here?	4.01 2.70	0.53 0.90

Table 4: Repetition chain for the construction *'might not be'* in dialogue SXWH, Spoken BNC, annotated with repetition index (RI), RI within dialogue turn (RI Turn), and distance from previous mention (Dist; in tokens).

than dialogue turns, since they more closely correspond to the temporal units used in previous work (e.g., Reitter et al., 2006a), and since the length of dialogue turns can vary significantly (see Table 2). To measure the surprisal of a construction c, we average over word-level surprisal values:

$$S(c;t,l) = \frac{1}{|c|} \sum_{w_i \in c} H(w_i|t,l) \qquad [3]$$

Surprisal estimates provide a computational approximation of the effort required to process a construction in context. We also measure the surprisal change (increase or reduction in processing effort) contributed by a construction c to its dialogue turn context, which we call the *facilitating effect* of a construction. The facilitating effect is positive when the construction has lower surprisal than its context, and negative when it has higher surprisal:

$$FE(c;t,l) = \log_2 \frac{\frac{1}{|s|-|c|} \sum_{w_j \in s, w_j \notin c} H(w_j|t,l)}{\frac{1}{|c|} \sum_{w_i \in c} H(w_i|t,l)} \qquad [4]$$

Due to human memory constraints, the facilitating effect of constructions is more likely to affect the processing of words that are produced immediately before and after the construction itself. We define the locus of the facilitating effect (s in Eq. 4) as the 10 tokens preceding and the 10 tokens following the construction.[8] The tokens exceeding the limits of the current dialogue turn are discarded. When the locus s corresponds to the construction itself, the facilitating effect equals 0.

6.2 Estimates of Surprisal

To produce surprisal estimates, we use a computational model of next word prediction which implements approximations of both the activation-based and the implicit learning mechanism: it is conditioned on local contextual cues while it learns from exposure to the global dialogue context. We use GPT-2 (Radford et al., 2019), a pre-trained autoregressive Transformer language model. We take GPT-2's attention mechanism (Vaswani et al., 2017) over the preceding context of a word as a proxy for the local activation-based mechanism: words in the more proximate dialogue context shape the model's expectations for next words, and thus their contextualised surprisal. As an implicit learning mechanism, we use the Transformer's standard learning rule, back-propagation on the cross-entropy next word prediction error, which has been successful at modelling a wide range of linguistic phenomena (Rumelhart and McClelland, 1986; Elman, 1991; Cleeremans and Elman, 1993; van Schijndel and Linzen, 2018). 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 in order to adapt it to the idiosyncrasies of spoken dialogic data.[9] We refer to this finetuned version as the *frozen* model. We use an attention window of length 50, i.e., the size of the local dialogue context, which may span over multiple dialogue turns (see Section 6.1).

Adaptive language model When estimating surprisal for a dialogue, we begin by processing the first turn using the frozen language model and then gradually update the model parameters after each turn, using back-propagation with cross-entropy loss. The magnitude of the learning rate is important for these updates to have the desired effect. The learning rate 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 learning rate, we randomly select 18 dialogues from

local priming 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.

[8]This is motivated by the fact that the average length of turns containing a construction is 31 tokens (median length is 21), with constructions being 3 to 7 tokens long—see Table 3.

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

the analysis split of the Spoken BNC[10] and run an 18-fold cross-validation for a set of six candidate learning rates: $1e-5$, $1e-4$, $\ldots$, 1. We fine-tune the model on each dialogue using one of these learning rates, and compute perplexity reduction 1) on the dialogue itself (*adaptation*) as well as 2) on the remaining 17 dialogues (*generalisation*). We select the learning rate yielding the best adaptation over cross-validaton folds ($1e-3$), while still improving the model's generalisation ability. See Appendix C.2 for further details.

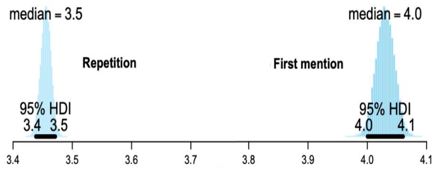
6.3 Statistical Modelling

To test **H1**, we split all occurrences of constructions by whether they are the first mention in a dialogue or a repetition. Our dataset consists of 8,562 first mentions and 24,541 repetitions. Using a Two Sample Bayesian t-test,[11] we compare the *S* distribution of first mentions to that of repetitions. We perform the same analysis for *FE* values.

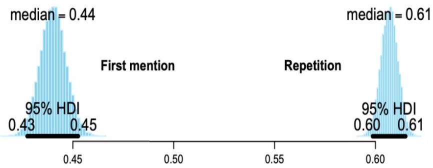
H2 and **H3** focus on analysing repetitions only. We label each occurrence with a *repetition index* (the first repetition of a construction has an index of 1, the second, 2, etc.), and with the *distance from the previous mention* in a dialogue, measured as the number of words between the first word of the current occurrence and the first word of the previous occurrence (see Table 4). We fit two linear mixed effect models using *S* and *FE* as response variables, and include multi-level random effects grouped by dialogue and individual speaker ID. To select the models' fixed effects, we start with a collection of motivated features—including repetition index and distance from previous mention—and perform an ablation selection procedure, iteratively removing features with the lowest significance, keeping only those that yield a p-value lower than 0.05.[12]

7 Results

We now present the results of our experiments, testing three hypotheses on the processing advantage (surprisal reduction and facilitating effect) of construction repetition. The final linear mixed effect models for both construction surprisal *S* and facilitating effect *FE* include repetition index and distance from the previous mention, which are directly related to our hypotheses, as well as construction length and repetition index within the current turn. The full specification of the best models, with fixed and random effect coefficients, is in Appendix D.



(a) Construction surprisal (*S*)



(b) Facilitating effect (*FE*)

Figure 1: Posterior predictive distributions for the mean *S* and *FE* according to the Bayesian t-test between first mentions and repetitions.

Repetition facilitates processing (H1) Figures 1a and 1b show that the posterior distributions of the mean *S* and *FE* do not overlap between groups. For both metrics, highest density intervals of difference between means do not include 0. In sum, we find surprisal of construction repetitions is lower than that of first mentions, and repetitions have a stronger facilitating effect than first mentions. Our first two predictions are thus confirmed.

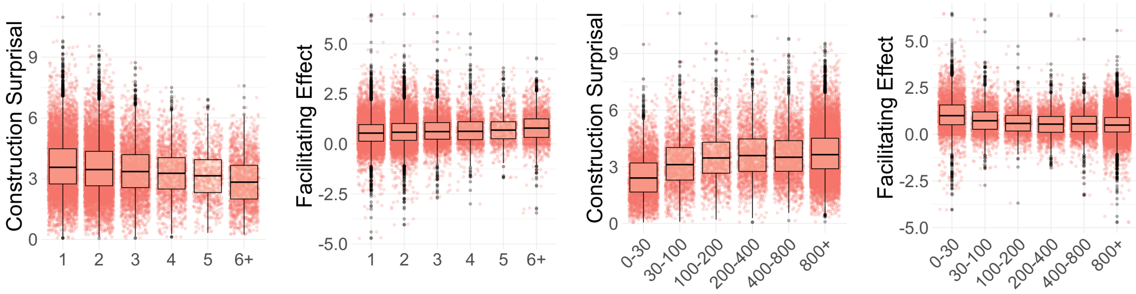
The processing advantage of repetition is cumulative (H2) The effect of repetition index is negative on *S* ($-24.85e-2, p < 2e-16$) and positive on *FE* ($7.57e-2, p < 2e-16$). Figures 2a and 2b show the opposite trajectories of the measures, with a stronger effect of repetition index on construction surprisal. In sum, we find that the surprisal of construction decreases, and their facilitating effect increases, as previous mentions accumulate. This confirms our second pair of predictions.

The processing advantage of repetition decays (H3) The distance of a construction from its previous mention has a positive effect on *S* ($9.66e-2, p < 2e-16$) and a negative effect on *FE* ($-4.29e-2, p < 2e-16$), also shown in Figures 2c and 2d. Surprisal increases, and facilitating effect decreases, as the current usage of a construction gets further away from its previous mention. Our third pair of predictions is thus confirmed.

[10]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.

[11]We use the t-test implemented in the 'Bayesian First Aid' R-JAGS package (`https://github.com/rasmusab/bayesian_first_aid`) with the default uninformative priors and a credible interval of 95%.

[12]The full list of features can be found in Appendix D.



(a) Repetition Index vs. *S* (b) Repetition Index vs. *FE* (c) Distance vs. *S* (d) Distance vs. *FE*

Figure 2: Construction surprisal (*S*, bits) and facilitating effect (*FE*) vs. repetition index and distance from previous mention (number of words). The first distance bin is the mean length of a turn containing a construction (Table 3).

8 Analysis

Having confirmed our hypotheses, we now further analyse the distribution of *FE* and *S* estimates, their relationship, and how their values across repetitions are influenced by additional factors.

8.1 Measures of Processing Advantage

Our first observation is that not only construction *repetition* but also construction *usage* comes with a processing advantage, as measured with both *S* and *FE*—a finding in line with prior work (e.g., Arnon and Snider, 2010; Bannard and Matthews, 2008; Tremblay et al., 2011; Janssen and Barber, 2012). On the one hand, as shown in Figure 1b, the posterior distribution of the mean *FE* spans over positive values for both first mentions and repetitions. The estimated mean *FE* of constructions is higher than the mean (0.07 ± 0.82) and median (0.01) *FE* of non-construction sequences in the Spoken BNC dialogues.[13] On the other hand, the posterior predictive mean value of *S* for constructions (Figure 1a) does not include the mean (5.59 ± 2.36) nor the median (5.36) *S* of non-construction sequences.

Our second observation is that the two metrics show similar but opposite patterns in our results. Based on the definition of the two metrics (Section 6.1)—these trends can be predicted a priori: it is more likely for a construction to have a facilitating effect if its surprisal is low; if construction surprisal is high, the context of the construction must be even more surprising for facilitating effect to occur. Empirically, we find that the Kendall's rank-correlation between facilitating effect and surprisal is -0.569 ($p < 2\text{e}-16$): 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 do not follow the predicted pattern. Some constructions have high surprisal and high facilitating effect:

A: So what have you got? what have you got going on with enrichments?
B: **I have to do** drama enrichment ($\boldsymbol{S = 5.46 \;\; FE = 1.32}$)

While there are cases where construction surprisal is low and facilitating effect is low or negative:[14]

A: But like I always really love strawberries but hate strawberry-flavoured things so I don't
B: I don't like strawberries **but I like** strawberry-flavoured things ($\boldsymbol{S = 2.24 \;\; FE = -0.70}$)

These examples show that our measures capture different types of context-dependent processing advantage.[15]

8.2 Other Predictors of Processing Advantage

Other factors that influence facilitating effect and surprisal beyond those directly related to our hypotheses are construction length and repetition index within a dialogue turn. Construction length has the strongest effect on both metrics (*S*: $-110.90\text{e}-2, p < 2\text{e}-16$; *FE*: $30.16\text{e}-2, p < 2\text{e}-16$): the longer the construction the lower its surprisal and the stronger its facilitating effect. Table 4 shows a full repetition chain for a construction of length 3; Table 5 (Appendix B) shows a chain for one of length 6. Because constructions, per se, have a processing advantage, and their repetition facilitates processing (see Section 7), construction repetition is more advantageous when constructions occupy

[13] We calculate *S* and *FE* of all 3- to 7-grams in 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 Section 8.2).

[14] A negative facilitating effect indicates that the surprisal of the construction is higher than the surprisal of its context.

[15] The examples have been selected among occurrences with *S* and *FE* higher or lower than the mean *S* / *FE* $\pm$ std.

a larger portion of processing time (which is proportional to the number of words).

The repetition index of a construction mention *within a dialogue turn* also has an effect on both metrics of processing advantage (*S*: $-29.48\text{e}-2, p < 0.05$; *FE*: $14.38\text{e}-2, p < 2\text{e}-16$); we find strong cumulativity effects for self-repetitions within the current dialogue turn.[16] Only 6.46% of the total construction occurrences have at least one previous mention in the same turn; yet when this is the case, the magnitude of *S* and *FE* increases with the number of previous local mentions. This interaction between cumulativity and recency (median distance between repetitions in the same turn is 7 words; across turns is 1208 words) indicates that processing advantage accumulates faster when repetitions are densely clustered. See Appendix E.

9 Conclusion

We have hypothesised that speakers repeat lexicalised constructions in dialogues because repetition eases information processing, and have formulated concrete predictions that follow from this hypothesis. To quantify the processing advantage of constructions we have proposed two surprisal-based measures, facilitating effect and construction surprisal, and have analysed how the values of these measures—estimated with a neural language model—vary as constructions are repeated. Although our experiments do not rely on direct measurements of the processing effort of human subjects, there is evidence that neural language models produce reliable estimates (Goodkind and Bicknell, 2018; Linzen, 2019; Schrimpf et al., 2021).

Our experiments on English spoken open domain dialogues confirmed our three predictions: (i) construction repetition reduces processing effort; (ii) the effort reduction increases with the frequency of repetitions and (iii) decreases with the distance between repetitions. These empirical results provide new evidence that construction repetition in dialogue is an efficient communication strategy. They thus complement prior work on the processing advantage of construction usage (cf. Section 2.1) and contribute to an understudied type of priming, with priming research traditionally focusing on repetitions of syntactic structures and lexical elements (cf. Section 1). Our findings reveal that the information processing efficiency of construction repetition results from a combination of the activation-based and implicit learning priming mechanisms. In line with activation-based accounts of priming, we find that the processing advantage of repetitions accumulates faster when repetitions are densely clustered, and it decays faster within more local distances. However, implicit learning is necessary to explain the fact that both cumulativity and decay effects are still present across distant repetitions. The discovered decreasing patterns of surprisal may seem to contradict the entropy rate constancy principle (Genzel and Charniak, 2002) and the principle of uniform information density (Jaeger and Levy, 2007), according to which surprisal remains stable over consecutive utterances. Yet we believe our findings can help explain these principles by providing insights into the information structure of individual utterances: the processing advantage of repeated constructions, which are not topic related, allows for progressively more information-dense topical and referential expressions. We conjecture that it is as a result of this balance that surprisal remains stable over utterances.

Besides contributing new empirical evidence on construction usage and repetition in dialogue, this study highlights the importance of a few key desiderata for the design of human-compatible computational dialogue models. First, models should both attend to the local dialogue context and use the global statistics collected throughout a dialogue for on-the-fly adaptation. This would have the natural effect of models being more likely to repeat constructions established as part of the dialogue lexicon. Second, although excessive and unnatural repetitions should be avoided in machine-generated utterances (Li et al., 2016; Holtzman et al., 2019), a certain degree of repetition makes a dialogue sound more natural. Human-like repetition patterns can be explicitly learned by auxiliary modules (Holtzman et al., 2018) or, as our study suggests, they may be implicitly acquired if next-word surprisal training and decoding objectives are complemented with context-dependent surprisal-based objectives. Simple techniques such as those proposed by Wei et al. (2021) and Meister et al. (2020) could be used to operationalise facilitating effect as a psycholinguistically motivated inductive bias to be used in training, and as a word choice criterion in decoding.

[16] The identity of the speaker producing previous mentions does not influence *FE* or *S*. All fixed effects related to speaker identity are discarded during the ablation procedure; see Section 6.3 and Appendix D.

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.

Kathryn Bock and Zenzi M Griffin. 2000. The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General*, 129(2):177.

Holly P Branigan, Martin J Pickering, and Alexandra A Cleland. 1999. Syntactic priming in written production: Evidence for rapid decay. *Psychonomic Bulletin & Review*, 6(4):635–640.

Holly P Branigan, Martin J Pickering, Andrew J Stewart, and Janet F McLean. 2000. Syntactic priming in spoken production: Linguistic and temporal interference. *Memory & Cognition*, 28(8):1297–1302.

Susan E Brennan. 1996. Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, 96:41–44.

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.

Axel Cleeremans and Jeffrey Elman. 1993. *Mechanisms of implicit learning: Connectionist models of sequence processing*. MIT press.

Alexandra A Cleland and Martin J Pickering. 2003. The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, 49(2):214–230.

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.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708.

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 C Frank. 2016. Investigating the sources of linguistic alignment in conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 526–536.

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.

Jeffrey L Elman. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, 7(2):195–225.

Alex B Fine and T Florian Jaeger. 2013. Evidence for implicit learning in syntactic comprehension. *Cognitive Science*, 37(3):578–591.

Alex B Fine and T Florian Jaeger. 2016. The role of verb repetition in cumulative structural priming in comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(9):1362.

Alex B Fine, T Florian Jaeger, Thomas A Farmer, and Ting Qian. 2013. Rapid expectation adaptation during syntactic comprehension. *PloS one*, 8(10):e77661.

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*.

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. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

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.

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.

Robert J Hartsuiker, Sarah Bernolet, Sofie Schoonbaert, Sara Speybroeck, and Dieter Vanderelst. 2008. Syntactic priming persists while the lexical boost decays: Evidence from written and spoken dialogue. *Journal of Memory and Language*, 58(2):214–238.

Patrick GT Healey, Matthew Purver, and Christine Howes. 2014. Divergence in dialogue. *PloS one*, 9(6):e98598.

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*.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1638–1649.

Christine Howes, Patrcik GT Healey, and Matthew Purver. 2010. Tracking lexical and syntactic alignment in conversation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

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.

T Florian Jaeger and Neal Snider. 2008. Implicit learning and syntactic persistence: Surprisal and cumulativity. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, volume 827812. Cognitive Science Society Austin, TX.

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*.

Michael P Kaschak, Timothy J Kutta, and Jacqueline M Coyle. 2014. Long and short term cumulative structural priming effects. *Language, cognition and neuroscience*, 29(6):728–743.

Michael P Kaschak, Renrick A Loney, and Kristin L Borreggine. 2006. Recent experience affects the strength of structural priming. *Cognition*, 99(3):B73–B82.

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.

Willem JM Levelt and Stephanie Kelter. 1982. Surface form and memory in question answering. *Cognitive Psychology*, 14(1):78–106.

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.

Tal Linzen. 2019. What can linguistics and deep learning contribute to each other? response to pater. *Language*, 95(1):e99–e108.

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

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185.

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.

Bill Noble and Raquel Fernández. 2015. Centre stage: How social network position shapes linguistic coordination. In *Proceedings of the 6th workshop on cognitive modeling and computational linguistics*, pages 29–38.

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

Martin J Pickering and Holly P Branigan. 1998. The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and language*, 39(4):633–651.

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. 2006a. 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, Frank Keller, and Johanna D Moore. 2011. A computational cognitive model of syntactic priming. *Cognitive science*, 35(4):587–637.

David Reitter, Johanna D Moore, and Frank Keller. 2006b. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.

DE Rumelhart and JL McClelland. 1986. On learning the past tenses of English verbs. In *Parallel distributed processing: explorations in the microstructure, vol. 2: psychological and biological models*, pages 216–271. MIT press Cambridge, MA.

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).

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.

G Underwood. 2004. The eyes have it. An eye-movement study into the processing of formulaic sequences.

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.

Jason Wei, Clara Meister, and Ryan Cotterell. 2021. A cognitive regularizer for language modeling. *arXiv preprint arXiv:2105.07144*.

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*.

Linda Wheeldon and Mark Smith. 2003. Phrase structure priming: A short-lived effect. *Language and Cognitive Processes*, 18(4):431–442.

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.

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 *ACL, system demonstration*.

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 to filter out topical and referential construction. 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*.

Table 5 shows a whole construction chain (from the first mention to the last repetition) for a construction of length 6.

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 4) 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

Speaker	RI	RI Turn	Dist	Turn	S	FE
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 [...]	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**	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	1.06	2.48

Table 5: A chain of repetitions of the construction *'I don't think you should be'* in dialogue S2AX of the Spoken BNC, annotated with repetition index (RI), repetition index within dialogue turn (RI Turn), and distance from previous mention (Dist; in tokens).

perplexity using Huggingface's finetuning script. We use early stopping over 5 epochs.[17] 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 perplexity is extremely high, and the perplexity trend against maximum sequence length is surprisingly upward. These two behaviours indicate that the pre-trained 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).

[17] 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.

C.2 Learning rate selection

To find the appropriate learning rate for on-the-fly adaptation (see Section 6.2), 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: $1e-5$, $1e-4$, $\ldots$, 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 6.1.

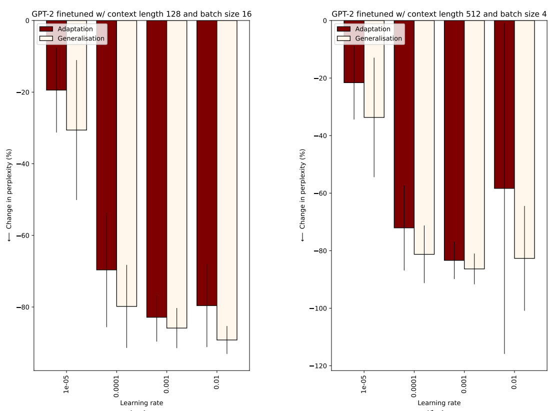
More precisely, for each dialogue $d \in D$, we calculate the perplexity of our two frozen models (Section C.1) on d and D d ($ppl_{before}(d)$ and $ppl_{before}(D)$, respectively). Then, we finetune the models on d using the six candidate learning rates, and measure again the perplexity over d and D d ($ppl_{after}(d)$ 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 $1e-3$ 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

Model	Learning rate	Max sequence length	Batch size	Best epoch	Perplexity finetuned	Perplexity pretrained
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.

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.



(a) (b)

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

D Linear Mixed Effect Models

As explained in Section 6.3 of the main paper, we fit linear mixed effect models using facilitating effect and construction surprisal as response variables and including multilevel random effects grouped by dialogues and individual speakers.[18] To select the fixed effects of the models, we start with a collection of motivated features and perform an ablation selection procedure, iteratively removing features with the lowest significance, and keeping only those that yield a p-value lower than 0.05. We start with the following features: the logarithm of the repetition index, the logarithm of the repetition index *within the current turn*, the logarithm of the distance[19] from the previous mention (computed in three ways: with respect to the previous mention of any speaker, of the current speaker, and of the other speaker), the logarithm of construction length (measures as the number of tokens in a construction), the logarithm of the number of tokens between the current occurrence and the first mention of a construction, and binary features indicating whether the previous mention is by the current speaker, whether it is produced by the initiator of the construction, whether the construction has been already uttered by both speakers, and whether the previous mention is in the current dialogue turn.

The ablation selection procedure yields two models with the following fixed effects: log repetition index, log repetition index within the current dialogue turn, log distance from the previous mention (of any speaker), and log construction length. The best model for facilitating effect is summarised in Listing 1 and the best model for construction surprisal in Listing 2.

E Local Effects of Processing Advantage

Table 7 shows the distribution of repetition indices within the dialogue turn. An index of n indicates that n previous mentions of the construction take place in the current dialogue turn. Figures 4a

[18]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.

[19]Distance is measured as the number of words between the first word of the current occurrence and the first word of the previous occurrence. We choose this strategy as there exist overlapping constructions and the distance values would be negative if we used the last word of the previous occurrence as a starting point to compute the distance.

Listing 1: Best linear mixed effect model for Facilitating Effect

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula:
logFE10 ~ 1 + logLength + logRepIndexInTurn + logRepetitionIndex +
    logDistance + (1 | `Dialogue ID`/Speaker)
   Data: data

REML criterion at convergence: 51869.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-7.3884 -0.6125 -0.0438  0.5574  8.4443

Random effects:
 Groups                Name        Variance Std.Dev.
 Speaker:`Dialogue ID` (Intercept) 0.006503 0.08064
 Dialogue ID           (Intercept) 0.006100 0.07810
 Residual                          0.478766 0.69193
Number of obs: 24540, groups:
Speaker:`Dialogue ID`, 364; Dialogue ID, 185

Fixed effects:
                    Estimate Std. Error         df t value Pr(>|t|)
(Intercept)        4.056e-01  5.335e-02  2.036e+04   7.603 3.02e-14
logLength          3.016e-01  2.901e-02  2.452e+04  10.394  < 2e-16
logRepIndexInTurn  1.438e-01  1.709e-02  2.451e+04   8.416  < 2e-16
logRepetitionIndex 7.569e-02  6.902e-03  2.360e+04  10.965  < 2e-16
logDistance       -4.290e-02  1.741e-03  2.309e+04 -24.638  < 2e-16

(Intercept)        ***
logLength          ***
logRepIndexInTurn  ***
logRepetitionIndex ***
logDistance        ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) lgLngt lgRIIT lgRptI
logLength   -0.909
lgRpIndxInT -0.177 -0.008
lgRpttnIndx -0.291  0.067 -0.031
logDistance -0.342  0.030  0.563  0.095
```

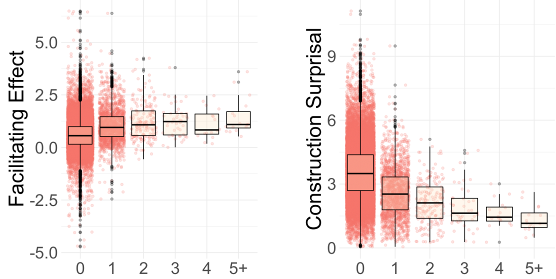
Listing 2: Best linear mixed effect model for Construction Surprisal

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: S ~ 1 + logLength + logRepIndexInTurn + logRepetitionIndex +
    logDistance + (1 | `Dialogue ID`/Speaker)
   Data: data

REML criterion at convergence: 78900.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.0885 -0.6807 -0.0779  0.6062  6.5359

Random effects:
 Groups                Name        Variance Std.Dev.
 Speaker:`Dialogue ID` (Intercept) 0.01282  0.1132
 Dialogue ID           (Intercept) 0.04292  0.2072
 Residual                          1.43852  1.1994
Number of obs: 24540, groups:
Speaker:`Dialogue ID`, 364; Dialogue ID, 185

Fixed effects:
                     Estimate Std. Error         df t value Pr(>|t|)
(Intercept)         4.866e+00  9.319e-02  1.810e+04  52.215   <2e-16
logLength          -1.109e+00  5.033e-02  2.451e+04 -22.042   <2e-16
logRepIndexInTurn  -2.948e-01  2.964e-02  2.452e+04  -9.943   <2e-16
logRepetitionIndex -2.485e-01  1.197e-02  2.346e+04 -20.761   <2e-16
logDistance         9.657e-02  3.028e-03  2.408e+04  31.889   <2e-16

(Intercept)        ***
logLength          ***
logRepIndexInTurn  ***
logRepetitionIndex ***
logDistance        ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) lgLngt lgRIIT lgRptI
logLength   -0.903
lgRpIndxInT -0.176 -0.007
lgRpttnIndx -0.289  0.068 -0.030
logDistance -0.339  0.031  0.563  0.096
```

	Previous mentions in the current dialogue turn								
Tot	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
33103	30965	1872	188	46	16	11	3	1	1

Table 7: The distribution of repetition indices *within the dialogue turn*.

and 4b show how facilitating effect and construction surprisal vary locally, for repetitions occurring within the same dialogue turn.



(a) (b)

Figure 4: Facilitating effect and construction surprisal (bits) against repetition index *within the current dialogue turn*.

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.