Is Information Density Uniform in Task-Oriented Dialogues?

Mario Giulianelli, Arabella Sinclair, Raquel Fernández
Institute for Logic, Language and Computation
University of Amsterdam
{m.giulianelli|a.j.sinclair|raquel.fernandez}@uva.nl

Abstract

The Uniform Information Density principle states that speakers plan their utterances to reduce fluctuations in the density of the information transmitted. In this paper, we test whether, and within which contextual units this principle holds in task-oriented dialogues. We show that there is evidence supporting the principle in written dialogues where participants play a cooperative reference game as well as in spoken dialogues involving instruction giving and following. Our study underlines the importance of identifying the relevant contextual components, showing that information content increases particularly within topically and referentially related contextual units.

1 Introduction

Due to production and perception errors, differences between individuals, and other sources of uncertainty, language use for information transmission can be thought to happen through a noisy channel. Effective and efficient information exchange under such conditions can be modelled using the tools of Information Theory (Shannon, 1948). Indeed, information-theoretic models have successfully accounted for surprisal in speech 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), providing psycholinguistic evidence that the information content of linguistic signals is related to comprehension processing effort.

Speakers, too, are sensitive to the properties of the communication channel. They are thought to simultaneously minimise their own production effort and the addressee's processing effort (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989). The most efficient way of dealing with both pressures, according to Information Theory, is to transmit information at a constant rate (Genzel and Charniak, 2002), making linguistic choices that reduce fluctuations in the density of the information transmitted. Evidence for the principle of uniform information density (UID; Jaeger and Levy, 2007; Jaeger, 2010) has been found at many levels of language production: speakers tend to reduce the duration of more predictable sounds (Aylett and Turk, 2004, 2006; Bell et al., 2003; Demberg et al., 2012); they tend to drop sentential material within more predictable scenarios (Jaeger and Levy, 2007; Jaeger, 2010; Frank and Jaeger, 2008); in spoken dialogue they are more likely to overlap at turn transitions when information density is low (Dethlefs et al., 2016); and the rate at which they transmit information in texts is uniform (Genzel and Charniak, 2002, 2003; Qian and Jaeger, 2011). Empirically, it is as yet unclear whether information density remains uniform throughout conversations (Vega and Ward, 2009; Doyle and Frank, 2015a,b; Xu and Reitter, 2018).

That information density is not always uniform in dialogue may be due to the complex structure of conversational context (Clark and Brennan, 1991), which not only includes previous utterances and world knowledge, but can also comprise preceding interactions between the interlocutors, their perceptual input, and their goals. This paper tests the UID principle on the previously unexplored setting of *task-oriented* dialogue, with its well-defined structural units and more constrained context than in open domain dialogues. To estimate information density, we use a pre-trained Transformer-based language model, which provides more robust measurements than the n-gram models used in prior work. We study whether, and within which structural units the UID principle holds, finding new evidence in support of it in certain structural units of both written cooperative reference games and spoken map navigation dialogues.[1] Our study highlights the importance of identifying relevant con-

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

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8271–8283
November 7–11, 2021.

textual structures, showing that topically and referentially related contextual units correspond to more uniform information transmission profiles.

2 Measuring Information Content

To investigate whether information density is uniform throughout a discourse, each lexical choice can be modelled as a random variable Y_i and its information density estimated as the Shannon information content $H(Y_i)$. For the UID principle to hold, the amount of information transmitted with every new word $H(Y_i)$ must remain constant. We can rewrite $H(Y_i)$ as $H(X_i|L_i, C_i)$, where C_i is the entire relevant context and L_i is the local context, both influencing lexical choice X_i. Typically, $H(X_i|C_i, L_i)$ is not estimated directly. The term is further decomposed into $H(X_i|L_i)$, the information content of X_i given the local context, and $I(X_i; C_i|L_i)$, the locally conditioned mutual information between X_i and the entire relevant context:

$$H(X_i|C_i, L_i) \equiv H(X_i|L_i) - I(X_i; C_i|L_i) \quad [1]$$

As the relevant context is built up, $I(X_i; C_i|L_i)$ is assumed to increase: next word prediction becomes easier when more contextual cues are available (Genzel and Charniak, 2002). So for the UID principle to hold—i.e., for $H(X_i|C_i, L_i)$ to remain constant in Eq. 1—the locally conditioned information content $H(X_i|L_i)$ must increase, too, as relevant context accumulates.

The local context of a word choice is typically taken to be the utterance or sentence, and these are also considered as the units of information transmission (Genzel and Charniak, 2002, 2003; Doyle and Frank, 2015a,b; Qian and Jaeger, 2011; Xu and Reitter, 2018). The information content of an utterance is computed by averaging over the negative logarithms of all locally contextualised word probabilities:

$$H(X|L) = -\frac{1}{|X|}\sum_{x_i \in X} \log_2 P(x_i|x_1, ..., x_{i-1}) \quad [2]$$

To remove the confounding effect of utterance length on information content (Keller, 2004), we use Xu and Reitter's (2018) normalised metric of utterance *information content*:

$$H'(X) = H(X)\frac{|L(n)|}{\sum_{S \in L(n)} H(S)} \quad [3]$$

where $L(n)$ is the set of all utterances of length n and $X \in L(n)$; for simplicity, we leave out the conditioning variable.

3 Data and Hypotheses

The UID principle is assumed to hold within a structural unit that determines the type and size of the overall relevant context C_i as used in Eq. 1. Genzel and Charniak (2002, 2003) show that, in texts, part of the relevant context is lexical (writers tend to reuse words that have already appeared in the discourse) and topically determined, as given by the paragraph structure of texts. In dialogue, defining a topically relevant contextual unit is not straightforward. Xu and Reitter (2018) use a topic segmentation algorithm to identify relevant units in unconstrained dialogues and show that information density is influenced by topic shift. Here we exploit the inherent (task-related) structure of task-oriented dialogues to test the UID principle within contextual units of different type and size. We analyse two corpora of task-oriented English dialogues: MapTask (MT, Anderson et al., 1991)[2] and PhotoBook (PB, Haber et al., 2019)[3]. Dialogue excerpts from both corpora can be found in Appendix A.

MT contains 128 transcribed spoken dialogues consisting of an instruction giver directing an instruction follower to navigate to a point on a map. The participants cannot see the other's map and their respective maps may contain slightly different landmarks. We consider two types of contextual unit: *a) the overall dialogue*: a series of landmarks are described in succession to help the instruction follower draw a path towards a goal location; *b) a dialogue transaction*: a dialogue excerpt related to reaching a certain landmark, manually annotated as part of the corpus. For both types of contextual unit, we also construct versions where we use the MT dialogue act annotation to filter out turns exclusively consisting of backchannels and other grounding acts (*'okay'*, *'mmhmm'*) common in spoken language.[4] This results in contextual units that focus on information-transmission dialogue acts and are more referentially coherent.

PB contains 2,500 dialogues where two participants without specified roles communicate via written chat. Each dialogue consists of 5 rounds: in each round, each participant sees a set of photographs which partially overlap with the set of images seen by their dialogue partner. The goal is to find out which images they have in common.

[2] http://groups.inf.ed.ac.uk/maptask

[3] https://dmg-photobook.github.io

[4] We exclude acknowledgements, attention and agreement checks, and pre-initiating moves.

The images available to each participant change in each round, but a subset reappears, thus triggering subsequent references to previously described photographs. This task design allows us to investigate the following types of contextual unit: *a) the overall dialogue*: throughout a game, all the photographs are about a certain domain (e.g., food or dogs); *b) a dialogue round*: different images are described in succession as participants try to figure out which ones they share in a given round; *c) an image reference chain*: the (non-adjacent) utterances that refer to a certain image across rounds (we use the automatic annotation of referring utterance chains by Takmaz et al., 2020).

We hypothesise that, in MT, the UID principle will be more visible at the transaction level, where the context is more topically coherent, than at the dialogue level, where a dozen different landmarks are brought up in succession—in particular when only information-transmission dialogue acts are taken into account. In PB, we expect the strongest effect to be present at the level of reference chains. Chains are determined both topically, by the target image, and lexically, by the conceptual pacts established in previous mentions of a target (Brennan and Clark, 1996). In rounds and dialogues, where several different images are described, topic and lexical choices are constrained by the image domain but the vocabulary used in previous turns is more varied. We thus expect the effect to be less pronounced at these two levels.

4 Modelling

To estimate the information content of an utterance we compute the log probabilities in Eq. 2 using GPT-2 (Radford et al., 2019), a pre-trained Transformer language model, which allows us to obtain more accurate probability estimates than n-gram models. We rely on HuggingFace's implementation of GPT-2 with default tokenizers and default parameters (Wolf et al., 2020). As GPT-2 was pretrained mainly on written text, it is less tuned to the idiosyncrasies of dialogue data. We therefore finetune it separately on a 70% split of each target corpus.[5] As shown in Table 1, finetuning yields a substantial reduction in the model's perplexity. More information on model parameters and the finetuning procedure can be found in Appendix B.

[5]We tried to finetune the language models on a combination of PhotoBook, MapTask and Spoken British National Corpus. The resulting model perplexities on PhotoBook and MapTask were higher than with the current approach.

	MapTask	PhotoBook	Penn Treebank
GPT-2 pre-trained	880.63	624.11	61.09
GPT-2 finetuned	48.36	41.79	43.39

Table 1: Word-level perplexity of the GPT-2 models on 30% held-out portions of the corpora.

We use the finetuned language models to estimate the information content (Eq. 3) of the 30% held-out portion of each corpus, and count turn positions (i.e., the positions of utterances within a dialogue—or a smaller structural unit) from the beginning of the relevant structural unit. [6] Following Xu and Reitter (2018), to test whether utterance information remains uniform we fit a linear mixed-effect model using the logarithm of information content as response variable and the logarithm of turn position as predictor. We include a random slope for the turn position and a random intercept term grouped by distinct dialogues, which allows us to model variation among individual speakers as a function of their addressee.

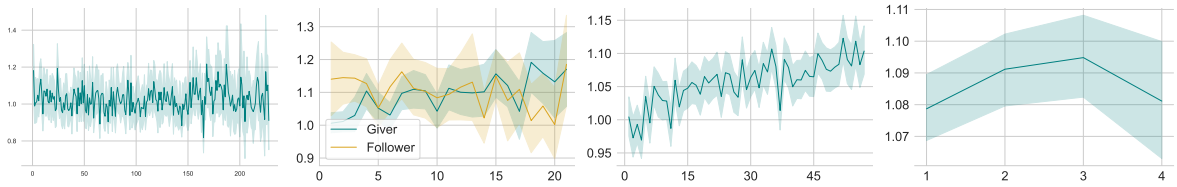
We adopt Genzel and Charniak's assumption that the mutual information $I(X_i; C_i|L_i)$ between an utterance and its context increases with turn position (Genzel and Charniak, 2002, see Section 2); so for $H(S_i|C_i, L_i)$ to remain stable, utterance information content $H(S_i|L_i)$, too, must increase. Consequently, we consider the UID principle to hold when turn position has a significant positive effect on information content.

Validation To validate our estimates of utterance information content, we replicate Genzel and Charniak's (2002; 2003) and Keller's (2004) study on the Wall Street Journal articles of the Penn Treebank (Mitchell et al., 1999)[7] using GPT-2 finetuned on this corpus (see Table 1). In the original studies, the authors measure the correlation between the position of sentences within newspaper articles—as well as within paragraphs—and the sentence information content, as measured using n-gram language models. As mentioned above, these studies assume that $I(X_i; C_i|L_i)$ increases as discourse context is built up, and test whether the locally conditioned information content $H(X_i|L_i)$, too, increases throughout articles and paragraphs.

In our validation study, we take both entire

[6]All dialogues, annotated with information content estimates, are provided in the supplementary material. Excerpts can be found in Appendix A.

[7]https://catalog.ldc.upenn.edu/LDC99T42; WSJ part of the corpus (sections 0–24).



(a) MT dialogues w/ bc (b) MT transactions w/o bc (c) PB dialogues (d) PB reference chains

Figure 1: Information content (y axis) against turn position (x axis) in MapTask (MT)—with or without backchannels (w/ bc and w/o bc, respectively)—and PhotoBook (PB) dialogues. Position is cut off at mean + 1sd, except for PB reference chains, where the median is 3 and the maximum position 4. Bootstrapped 95% confidence bands.

articles and paragraphs as structural units and count sentence positions from the beginning of the relevant unit. Our linear mixed-effect models show a significant positive effect of sentence position on information content both within articles ($\beta = 1.65\text{e}{-2}, p < 0.001$) and within paragraphs ($\beta = 1.53\text{e}{-2}, p < 0.01$). To reproduce the original experimental setting, we further train an n-gram language model with interpolated Kneser-Ney smoothing using Keller's (2004) data split and select the configuration with the lowest perplexity on the test set, a 3-gram model with a discount value of 0.8. In line with previous work, we find a positive Kendall's rank-correlation[8] between sentence position and information, as measured with the n-gram model as well as with the Transformers. The original results are therefore replicated.[9]

5 Results

We test whether the UID principle holds in MT and PB using the procedure presented in Section 4. The full results of our statistical analysis can be found in Appendix E (Tables 6 and 7). Recall that for the principle to hold, the locally conditioned information content $H(X_i|L_i)$ must increase with the position of X_i in the relevant context unit C_i. The local context L_i is defined as a dialogue turn.

5.1 MapTask

When we take entire MT dialogues as the contextual unit, we do not find a positive effect of turn position on information content, regardless of whether we focus in information-transmission dialogue acts (see Figure 1a for the results with all dialogue acts). In contrast, the types of dialogue act considered affect our results on transactions. We fail to find an effect in transactions with backchannels but the linear mixed-effect models show a positive effect of turn position within transactions without backchannels ($\beta = 2.38\text{e}{-2}, p < 0.001$). We attribute these findings to the nature of the task. Over the course of a dialogue, speakers traverse a map naming different landscape features and therefore are unable to establish more than a minimal level of linguistic routine at the dialogue level. Transactions, on the other hand, correspond to more referentially constrained subtasks; this becomes more evident when information-transmission dialogue acts are isolated from transmission-coordination acts. Analysing the instruction giver and follower information-transmission turns independently reveals that there is no significant effect for instruction followers; the overall positive effect is driven by the instruction givers ($\beta = 3.46\text{e}{-2}, p < 0.001$; see Figure 1b). This reflects the asymmetric nature of information transmission in MT dialogues.

5.2 PhotoBook

The effect of position on information content is positive within the PB dialogues ($\beta = 3.13\text{e}{-2}, p < 0.001$); Figure 1c shows a consistently increasing sawtooth pattern for information content, providing evidence that participants optimise their information-transmission strategy throughout PB games. Information content slightly decreases within game rounds ($\beta = -0.74\text{e}{-2}, p < 0.005$), yet this effect is mainly due to the higher estimates obtained for the first turns of these contextual units (see Figure 3e in Appendix E), often used by participants to coordinate on how to start the new round. Because multiple images are discussed in a round, this contextual unit seems not to capture the relevant context of individual dialogue turns nor be

[8] Our data consist of multiple measurements for each sentence position (one for each document), thus causing a large number of ties (i.e., multiple entries with the same sentence position but different entropy estimates). We choose Kendall's test for all our experiments because it deals with ties better than other correlation tests such as Spearman's or Pearson's.

[9] A detailed description of the experimental setup and the full results can be found in Appendix D.

large enough to display the participants' overall information transmission strategy that we observe at the dialogue level.

Finally, as hypothesised, the effect of position on information content is positive at the reference chain level ($\beta = 1.27\text{e}{-2}, p < 0.001$). As participants re-refer to an image over the game, they increase the density of their messages (as shown in Figure 1d) and also decrease message length (Kendall's correlation between position in chain and length is $\tau = -0.268, p < 0.001$). Thus, as reference chains unfold, the *reduction* process observed by Takmaz et al. (2020) is complemented by *information compression*. The relatively low magnitude of the fixed effect as well as that of the correlation between utterance length and chain position, however, suggest that the process we see at play is not only one of compression and reduction. Figure 1d indeed shows that the fourth position in a chain often comes with a decrease in information content, perhaps indicating that once a conceptual pact has been established between interlocutors, referential expressions can be significantly simplified without losing referential power—as in the following reference chain (information content estimates in parenthesis):

1. *'Man eating slice of pizza'* (0.69)
2. *'last one for me is guy with pizza'* (0.78)
3. *'pizza eater'* (0.91)
4. *'pizza'* (0.67)

6 Conclusion

We investigated to what extent the principle of uniform information density holds in two corpora of English task-oriented dialogues. We have related the properties of task-determined contextual units to patterns of information transmission and have hypothesised that the UID principle holds to a stronger degree in more topically coherent and reference-specific contextual units.

Our hypotheses are confirmed in PhotoBook, where we find evidence that dialogue participants use rational strategies of information transmission over an entire dialogue. We do not observe uniformity of information in the MapTask dialogues and transactions as a whole, similarly to other negative results in interactive settings (e.g., Vega and Ward, 2009; Doyle and Frank, 2015b). Yet the effect is present within MapTask transactions when we restrict our analysis to information-transmission dialogue acts: these make for a more topically and referentially coherent contextual unit. Indeed, the organisation of context can be complex in dialogues. We have shown that theoretically motivated contextual units such as reference chains in PhotoBook and information-transmission acts in MapTask transactions are good candidates to characterise the relevant context over which participants deploy strategies of information compression.

We are aware that the assumptions used to test the UID principle, which we have adopted from Genzel and Charniak's seminal work (2002)—i.e., that context informativeness increases as strongly as sentence entropy as discourse is built up—can be controversial. Nevertheless, in this paper we have followed this line of reasoning, used in previous work (Genzel and Charniak, 2003; Vega and Ward, 2009; Qian and Jaeger, 2011; Doyle and Frank, 2015b; Xu and Reitter, 2018), and applied it to novel data and contextual units. In Giulianelli and Fernández (2021), we go one step further and empirically test these assumptions for the first time, using direct estimates of the contextualised entropy $H(S_i|C_i, L_i)$ of an utterance and thus of the informativity of its linguistic context $I(S_i; C_i, L_i)$.

The study presented in this paper provides new empirical evidence on language production in dialogue which we believe can directly inform the development of natural language generation models. Our findings suggest that models that take relevant contextual units into account (Takmaz et al., 2020; Hawkins et al., 2020) are better suited for reproducing human patterns of information transmission, and confirm that the use of training objectives that enforce a uniform organisation of information density (Meister et al., 2020; Wei et al., 2021) is a promising avenue for training language models.

Acknowledgements

We would like to thank Jaap Jumelet for a helpful discussion on neural language models, the anonymous EMNLP-2021 reviewers for their valuable comments, as well as the anonymous ACL-2021 reviewers for feedback that led to a considerable improvement of the first version of this paper. 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

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC Map Task corpus. *Language and Speech*, 34(4):351–366.

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.

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.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

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.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association.

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.

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.

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*.

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.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910.

Robert Hawkins, Minae Kwon, Dorsa Sadigh, and Noah Goodman. 2020. Continual adaptation for efficient machine communication. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 408–419.

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.

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.

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.

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.

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, Online. Association for Computational Linguistics.

Marcus P. Mitchell, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3 LDC99T42 Web Download. Linguistic Data Consortium.

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.

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

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, Reuse, Reduce: Generating subsequent references in visual and conversational contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.

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.

Jason Wei, Clara Meister, and Ryan Cotterell. 2021. A Cognitive Regularizer for Language Modeling. *CoRR*, abs/2105.07144.

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.

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

Appendix

A Dialogue Excerpts

Tables 2 and 3 show excerpts of MapTask and PhotoBook dialogues. The dialogues are annotated with turn positions (within different contextual units), speaker identifier, and information content estimates. The speaker identifiers in MapTask refer to the speaker roles of instruction givers (G) and followers (F).

B Transformer Language Models

We experiment with GPT-2 (Radford et al., 2019), an autoregressive Transformer-based (Vaswani et al., 2017) language model, relying on HuggingFace's implementation with default tokenizers and default parameters (Wolf et al., 2020).[10] The maximum sequence length is set equal to the maximum utterance length in the corpus: 320 for Penn Treebank, 150 for MapTask, and 40 for PhotoBook. As the pre-trained model yields high perplexity on the dialogue corpora (Table 1), we finetune[11] it on 70% of each target corpus and leave out 30% of the dataset to compute the model's evaluation perplexity and to conduct our statistical analysis. The training and held-out portions of PhotoBook

[10] The pre-trained model is named `gpt2` in HuggingFace.

[11] We use HuggingFace's finetuning script https://github.com/huggingface/transformers/blob/master/examples/pytorch/language-modeling/run_clm.py.

Dialogue position	Transact. position	Transact. number	Speaker	Utterance	Information content
1	1	1	G	have you got the missionary camp at the right-hand side	0.68
2	2	1	F	yeah uh-huh	0.70
3	3	1	G	okay	0.68
4	4	1	G	have you got the start	0.83
5	5	1	F	yeah i've got the start	0.87
6	6	1	G	is it at the top of it	0.75
7	7	1	F	uh-huh	0.50
8	8	1	G	right okay	0.89
9	9	1	G	go to the right i think it is yeah right go to the right about two centi– about two three centimetres	0.93
10	10	1	F	right	0.87
11	1	2	G	go down two three centimetres	1.08
12	2	2	F	down	1.46
13	3	2	F	towards the missionary camp	0.73
14	4	2	G	uh-huh	0.50
15	5	2	G	you're at the top and that	0.89
16	6	2	F	uh-huh	0.50
17	1	3	G	go right about two go right three centimetres yeah	1.21
18	2	3	F	uh-huh	0.50
19	3	3	F	past the missionary camp	1.49
20	4	3	G	past missionary camp	2.24

Table 2: The first three transactions of a MapTask dialogue (dialogue id: q6nc5).

Dialogue position	Round position	Round number	Speaker	Utterance	Information content
1	1	1	B	dog in hot dog costume	0.79
2	2	1	A	dont have that	0.65
3	3	1	B	pug looking at cow	1.19
4	4	1	B	out a window	1.86
5	5	1	A	i have the pug	0.82
6	6	1	B	dog on persons lap, person watching tv, dog looking at camera. yellow walls.	1.39
7	7	1	A	dont have that	0.65
8	1	2	B	hot dog	1.13
9	2	2	A	got it	1.09
10	3	2	A	pug looking out the window	1.15
11	4	2	B	do not have	0.83
12	5	2	B	dog on persons lap, person watching tv, dog looking at camera. yellow walls.	1.55
13	6	2	A	got that	1.19
14	7	2	B	wait sorry i don't have that as a main one my bad lol	1.53
15	8	2	B	pug on plaid pants lap	1.66
16	9	2	A	dont have that	0.65
17	10	2	A	i already submitted the yellow one though	1.53
18	11	2	B	two guys chain link fence behind them two dogs on chess table	1.99
19	12	2	A	nope	0.62
20	13	2	B	i do have it just not as a main so ur good	1.78

Table 3: The first two rounds of a PhotoBook dialogue (dialogue id: 1770).

consist of games 0-1751 and 1752-2501 respectively; the training and held-out set of MapTask comprise dialogues *q1ec1-q6nc2* and *q6nc3-q8nc8*. One version of GPT-2 is finetuned for 30 epochs on PhotoBook dialogues with a learning rate of $5e-05$ and batches of size 64; a second version is finetuned for 60 epochs on MapTask dialogues with a learning rate of $1e-05$ and batches of size 16; the last version is finetuned for 30 epochs on Penn Treebank articles with a learning rate of $5e-05$ and batches of size 8. The other finetuning parameters are set to their default values.

Utterance beginning and end are used as context cues but their information content is not computed. Furthermore, for the dialogue corpora, we try prepending input utterances with dialogue turn cues ("A: ", "B: ") as a hint to the language models that the data is conversational; the information content of these speaker identifying tokens is never computed. This modification of the input text does not consistently reduce the models' perplexity scores. The perplexity of the pre-trained and finetuned models on the target corpora is reported in the main paper.

C Effects of Finetuning

The following are the main effects of finetuning GPT-2 on MapTask dialogues:

- GPT-2 finetuned on MapTask assigns lower perplexity to disfluencies. While the pre-trained model assigns high information content to utterances that contain disfluencies, this is not the case for the finetuned model.
- Backchannels also become less surprising with finetuning: the information content of, e.g., *okay*, *mmhmm*, *well*, *right*, *erm*, *yeah*, *no*, *aye* decreases by 25% to 75
- With finetuning, GPT-2 doesn't only get used to features of transcribed speech: expressions that refer to MapTask landmarks also become more likely (e.g., *the rapids*, *a rope bridge*, *the gold mine*)
- Simple spatial indications (*towards the bottom left-hand corner*, *on the left-hand side*) are among the utterances with the lowest surprisal.

These are the main effects of finetuning GPT-2 on PhotoBook dialogues:

- The finetuned model is less surprised by utterances types that are frequent in the corpus; the least surprising expressions are *I have*, *I don't have that one*, *I don't have that*, *No, I don't have that one*. For the pre-trained model, on the other hand, the least surprising expressions are more generic: *No, I don't think so*, *I'm not sure*, *I don't*, *What do you think?*.
- Among the most surprising utterances for the pre-trained model are some that are specific to PhotoBook games: *submit bye*, *loading may be frozen*. For these two utterances, e.g., surprisal decreases by 1/4 and 1/3 respectively after finetuning.
- Written chat language becomes less surprising: e.g., the surprisal for *kk done* decreases by one third.
- Utterances at first dialogue positions become in general less surprising (cf. initial drop in the PhotoBook dialogue graph) but the decrease in surprisal for greetings is not always very substantial: e.g., the surprisal for *hi* and *hey there* decrease by one third and one seventh respectively.

D Penn Treebank Replication Study

We use the Wall Street Journal part of the Penn Treebank, divided into a training set (section 0–20) and a test set (sections 21–24). The training set contains 41,128 sentences (Keller (2004) reports 42,075 sentences), the test set 8,594 (Keller (2004) reports 7,133). Each article is treated as a separate text and sentence positions are computed by counting sentences from the beginning of the article. The sentence positions in the test set varied between 1 and 118 (Keller (2004) reports 1-149). The n-gram probabilities were computed by Keller (2004) using a language model with smoothing by absolute discounting, whereas Genzel and Charniak (2002) do not report the specifics of their language model. We rely on NLTK's implementation of an n-gram language model with interpolated Kneser-Ney smoothing (Bird et al., 2009). Sentence beginning and sentence end are used as context cues but their information content never computed. We train language models with $n \in (2, 3, 4, 5)$ and with discount values $n \in (0.1, 0.2, \ldots, 0.9)$ on the training set and select the language model with the lowest perplexity on the test set. We do not split

	Cut-off = 25	Cut-off = 76	Cut-off = ∞
Raw data	τ	τ	τ
3-gram (Keller, 2004)	0.060^{**}	0.081^{**}	0.071^{**}
3-gram (ours)	0.076^{**}	0.081^{**}	0.079^{**}
GPT-2 pre-trained	0.032^{**}	0.055^{**}	0.054^{**}
GPT-2 finetuned	0.070^{**}	0.080^{**}	0.080^{**}
Binned data	τ	τ	τ
3-gram (Keller, 2004)	0.639^{**}	0.243^{**}	0.135
3-gram (ours)	0.733^{**}	0.109	0.118
GPT-2 pre-trained	0.533^{**}	0.512^{**}	0.077
GPT-2 finetuned	0.693^{**}	0.387^{**}	0.119

Table 4: Kendall's rank-correlation between sentence information and sentence position for the Penn Treebank test set. Significance: '**' $p < 0.001$, '*' $p < 0.01$, '' $p \geq 0.05$.

	Cut-off = 25	Cut-off = 76	Cut-off = ∞
Raw data	τ	τ	τ
3-gram (Keller, 2004)	0.078^{**}	0.093^{**}	0.081^{**}
3-gram (ours)	0.082^{**}	0.087^{**}	0.087^{**}
GPT-2 pre-trained	0.034^{**}	0.054^{**}	0.054^{**}
GPT-2 finetuned	0.077^{**}	0.084^{**}	0.084^{**}
Binned data	τ	τ	τ
3-gram (Keller, 2004)	0.671^{**}	0.147	0.170^{**}
3-gram (ours)	0.740^{**}	0.099	0.097
GPT-2 pre-trained	0.453^{*}	0.448^{**}	0.101
GPT-2 finetuned	0.680^{**}	0.347^{**}	0.104

Table 5: Kendall's rank-correlation between sentence information and sentence position, with sentence length partialled out, for the Penn Treebank test set. Significance: '**' $p < 0.001$, '*' $p < 0.01$, '' $p \geq 0.05$.

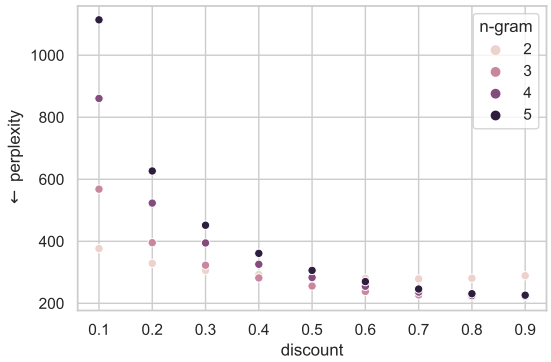


Figure 2: Perplexity on Penn Treebank test set obtained by n-gram language models with Kneser-Ney smoothing and interpolation.

the data into multiple test tests (Genzel and Charniak, 2002, 2003) as this was shown not to alter the sentence information estimates (Keller, 2004; Xu and Reitter, 2018). The best language model is the 3-gram model with a discount value of 0.8, which achieves a perplexity of 335.80 on the test set. The perplexity obtained using NLTK's evaluation script is 221.57 (Figure 2) as it is calculated by taking into account beginning and end of sentence symbols.

We use the n-gram language model as well as the GPT-2 language model (as described in Section B) to estimate the information content of all sentences in the test set and measure the correlation with sentence position. In Genzel and Charniak's (2002) original work, the correlation between sentence position and sentence information is computed by binning the sentence information data points based on their sentence position. Correlation is measured between sentence position indices 1-25 and the average sentence information estimated for the respective sentence position. Keller (2004) also measures the raw correlation between all sentence position-information pairs, without binning. Neither work reports the correlation measure used. We use Kendall's rank-correlation as it is less sensitive than Spearman's rank-correlation to the large amount of ties (position-information pairs with the same position index) in our data. Moreover, whereas Genzel and Charniak (2002) select a single sentence position cut-off ($c = 25$), in Keller's (2004) study three variants of the cut-off are used ($c = 25$, $c = 76$, and no cut-off). We also compute correlation at these three levels. Finally, following Keller (2004), we compute the partial correlation between sentence position and sentence information, excluding the effect of sentence length. The results are reported in Tables 4 and 5.

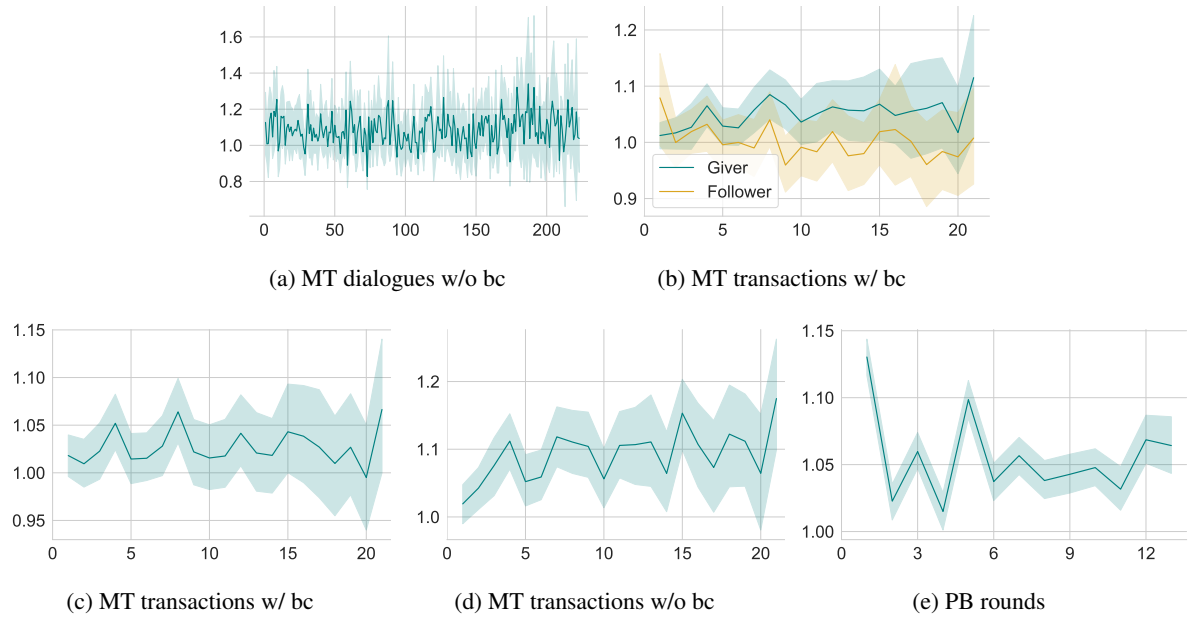
E Experimental Results

Tables 6 and 7 summarise the results of our statistical analysis, as introduced in Section 4. In both tables, the logarithm of information content is the response variable and the logarithm of turn position is the fixed effect. We include a random intercept grouped by distinct dialogues and a random slope for the turn position. Fixed effects with significant coefficient estimates are marked in bold. The Random effects columns show the standard deviation of the random effects (Coeff.) and the residual standard deviation. *The UID principle is considered to hold when turn position has a significant positive effect on information content.*

Figure 3 shows the patterns of information content against turn position for the contextual units whose patterns are not displayed in Section 5.

F Computing Infrastructure

The models were trained and evaluated on a computer cluster with Debian Linux OS. Parallelization over four GPUs was implemented for the finetuning of GPT-2. All information content computations were executed using used a single GPU. The GPU nodes are GPU GeForce 1080Ti, 11GB GDDR5X, with NVIDIA driver version 418.56 and CUDA version 10.1.



(a) MT dialogues w/o bc (b) MT transactions w/ bc

(c) MT transactions w/ bc (d) MT transactions w/o bc (e) PB rounds

Figure 3: Information content (y axis) against turn position (x axis) in MapTask (MT) dialogues and transactions—with or without backchannels (w/ bc and w/o bc, respectively)—and PhotoBook (PB) dialogue rounds. Position is cut off at mean + 1sd. Bootstrapped 95% confidence bands.

		Fixed effects			Random effects (Std. Dev.)	
		Estimate	**Std. Error**	**Pr(>|t|)**	**Coeff.**	**Residual**
MT dialogues w/ bc	Intercept	0.07e−2	2.40e−2	0.98	11.44e−2	30.05e−2
	Position	−0.70e−2	0.37e−2	0.06	0.80e−2	
MT dialogues w/o bc	Intercept	3.37e−2	3.20e−2	0.30	15.05e−2	26.83e−2
	Position	0.03e−2	0.57e−2	0.96	1.98e−2	
MT transactions w/ bc	Intercept	−2.95e−2	1.50e−2	0.06	7.90e−2	30.07e−2
	Position	0.03e−2	0.37e−2	0.93	0.34e−2	
	Intercept (givers)	−2.20e−2	1.49e−2	0.15	7.38e−2	28.40e−2
	Position (givers)	0.92e−2	0.46e−2	0.06	0.90e−2	
	Intercept (followers)	−5.01e−2	2.30e−2	0.04	10.88e−2	31.40e−2
	Position (followers)	0.60e−2	0.71e−2	0.41	1.50e−2	
MT transactions w/o bc	Intercept	−0.93e−2	1.61e−2	0.57	7.98e−2	26.80e−2
	Position	2.38e−2	0.49e−2	< 0.01	0.96e−2	
	Intercept (givers)	−3.78e−2	1.70e−2	0.03	8.20e−2	25.47e−2
	Position (givers)	3.46e−2	0.53e−2	< 0.01	0.19e−2	
	Intercept (followers)	9.04e−2	3.10e−2	< 0.01	13.10e−2	28.27e−2
	Position (followers)	−1.30e−2	1.38e−2	0.36	5.50e−2	

Table 6: Results of linear mixed-effect models on the MapTask data.

		Fixed effects			Random effects (Std. Dev.)	
		Estimate	**Std. Error**	**Pr(>|t|)**	**Coeff.**	**Residual**
PB dialogues	Intercept	−12.21e−2	0.90e−2	< 0.01	17.52e−2	37.66e−2
	Position	3.13e−2	0.24e−2	<0.01	3.77e−2	
PB rounds	Intercept	−0.99e−2	0.70e−2	0.16	15.14e−2	37.82e−2
	Position	−0.73e−2	0.26e−2	<0.01	3.46e−2	
PB chains	Intercept	−5.92e−2	0.36e−2	<0.01	14.86e−2	28.95e−2
	Position	1.27e−2	0.27e−2	<0.01	4.70e−2	

Table 7: Results of linear mixed-effect models on the PhotoBook data.