Investigating perception of spoken dialogue acceptability through surprisal

Citation for published version:
Wallbridge, S, Bell, P & Lai, C 2022, Investigating perception of spoken dialogue acceptability through surprisal. in H Ko & JHL Hansen (eds), *Proceedings of Interspeech 2022.* ISCA, pp. 4506-4510, Interspeech 2022, Incheon, Korea, Democratic People's Republic of, 18/09/22. https://doi.org/10.21437/Interspeech.2022-10808

Digital Object Identifier (DOI):
10.21437/Interspeech.2022-10808

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of Interspeech 2022

Investigating perception of spoken dialogue acceptability through surprisal

Sarenne Wallbridge, Peter Bell, Catherine Lai

Centre for Speech Technology Research, University of Edinburgh, United Kingdom

{s1301730, peter.bell, c.lai}@ed.ac.uk

Abstract

Surprisal is used throughout computational psycholinguistics to model a range of language processing behaviour. There is growing evidence that language model (LM) estimates of surprisal correlate with human performance on a range of written language comprehension tasks.

Although communicative interaction is arguably the primary form of language use, most studies of surprisal are based on monological, written data. Towards the goal of understanding perception in spontaneous, natural language, we present an exploratory investigation into *whether the relationship between human comprehension behaviour and LM-estimated surprisal holds when applied to dialogue*, considering both written dialogue, and the lexical component of spoken dialogue. We use a novel judgement task of dialogue utterance acceptability to ask two questions: "How well can people make predictions about written dialogue and transcripts of spoken dialogue?" and "Does surprisal correlate with these acceptability judgements?".

We demonstrate that people can make accurate predictions about upcoming dialogue and that their ability differs between spoken transcripts and written conversation. We investigate the relationship between global and local operationalisations of surprisal and human acceptability judgements, finding a combination of both to provide the most predictive power.

Index Terms: psycholinguistics, spoken dialogue, speech perception, discourse structure

1. Introduction

Recent developments in automatic language modelling have made it possible to test (and extend) psycholinguistic models of human language comprehension. In particular, autoregressive Language Model (LM) estimates of surprisal have been found to correlate with aspects of language perception, including reading times and grammatical acceptability judgements of written text. This is thought to be caused, at least in part, by the shared use of predictive processing that both people and such LMs rely on [1, 2]. In this work, we explore whether improved language modelling capabilities allow us to explore perception of more natural forms of language: interactive communication.

Although monological texts have been the primary testing ground for psycholinguistic theories of language comprehension, the cognitive mechanisms for comprehension are tuned to natural, spontaneous language [3]. To better understand comprehension, we must study perception of more realistic language. Controlled linguistic stimuli differ from casual language in numerous ways. In particular, they are often isolated [4]. Accounting for context above sentence-level has been a major hurdle in modelling realistic perception, but modern LMs allow integration of much longer contexts than previously possible [5, 6].

Speech is perhaps the most intrinsic modality for communication, but progress in developing models of perceptual salience for the speech signal has been markedly slower than for its textual counterpart, with good reason. Although we learn to use and understand spoken language long before learning to read or write, it differs from text in a number of ways that make modelling more complex. One feature of particular interest here is that speech signals are generated using multiple channels of information transmission: the lexical channel of written language (*which* words are used), and an additional non-lexical channel (*how* those words are said) [7]. Channel access changes the way in which we design communicative signals. In this paper, we ask whether surprisal still aligns with perception in spoken transcripts.

As a step towards modelling language perception in more realistic communicative settings, we examine perception though a novel acceptability rating task over dialogue turns. The task is designed to be applied to both written dialogues and transcripts of spoken dialogues, and investigates differences in how lexical information is distributed between these modalities during communicative interaction. Following works that have demonstrated a relationship between LM surprisal estimates and human comprehension behaviour on written, monological data, we explore the relationship between different definitions of turn-level surprisal and human judgements of dialogue.

2. Background

2.1. Surprisal and Language Comprehension

Human language comprehension is often formalised, at least in part, as a predictive process. Surprisal Theory, one of the most widely-adopted theories of human language comprehension, suggests that the cognitive cost of processing a linguistic segment is determined by how predictable the segment is in its preceding context [8, 9]. It draws on the information-theoretic formalisation of surprisal, which quantifies the amount of information conveyed by a unit as the uncertainty associated with its occurrence [10]. The standard definition of conditional surprisal is the negative log-probability of a unit in $[u_1, ..., u_N]$ conditioned on its prior context:

$$S(u_n) = -\log_2 p(u_n|u_{<n}). \qquad (1)$$

Surprisal theory has been used to model a range of human language comprehension behaviour including processing of syntactic and pronoun ambiguity, sentence interpretation, and word predictability effects as measured by self-paced reading times and eye-tracking studies [11, 12, 13, 14]. As such, we use surprisal theory as a basis for investigating perception of both written dialogue, and transcripts of spoken dialogue.

2.2. Surprisal and Language models

Language models have been inextricably linked to surprisal since their inception. LMs estimate the probability of a word in context – its *predictability*. Recent advances in the capability of LMs to capture longer contexts has prompted the use of LMs to study language and comprehension behaviour. The primary focus of such prior works has been sentence comprehension: investigating the relationship between LM surprisal estimates and self-paced reading times, gaze duration [13, 14, 15], acceptability judgements [16, 17, 18], and brain response data [11, 1].

Though there is general consensus that a relationship between LM surprisal estimates and human perception exists, there are still important aspects that require investigation. Architectural differences in LMs have been found to influence aspects of psychometric predictive power differently [1]. For example, [15] find that surprisal estimates from BERT are highly predictive of acceptability judgments, yet remarkably poor for reading time estimates, and [13] demonstrates that once perplexity is controlled for, syntactic generalization is largely determined by model architecture. The relationship between LM quality and psychometric predictive power doesn't necessarily generalise to typologically different languages [19]. Sentence processing behaviour is also affected by other linguistic features independently of surprisal, including local statistics such as word n-gram frequency [20, 21].

The predictive power of surprisal estimates has been explored in other aspects of perception, e.g., essay quality [22], but there is far less evidence for the relationship between surprisal and perception beyond the sentence processing task.

2.3. Extending surprisal to (spoken) dialogues

The vast majority of computational psycholinguistic theories have been developed using monologues. However the most fundamental forms of language-use are interactive [3]. Theories of communication are often centered around interaction and collaboration, e.g., as a joint process where interlocutors collaborate to build common ground [23]. The interactive nature of dialogue likely requires expectations to be conditioned on additional pragmatic features and wider discourse context [24, 25].

Automatic language modelling has been similarly focused on monologue data. Recent work has begun to explore learning latent spaces that are more suited to dialogue by augmenting training objectives to encode dialogue-specific structure and amplify the importance of temporal dependencies between utterances [26, 27, 28, 29, 30]. However, it is unclear whether these strategies encourage encoding of the interactive, joint nature of communication emphasized by psycholinguistic theories.

Although closely related, spoken and written communication are generated in fundamentally different conditions. The additional non-lexical channel of speech conveys novel information in its own right [31, 32, 33] and interact with the lexical channel to mark novel content, disambiguate lexical information [34, 35], and moderate the distribution of information during communication [7]. Incrementality also asserts much stronger pressure in the spoken domain where utterance design in dialog is often modelled as a parallel and predictive process [36]. As such, lexical information is likely to be distributed differently across written and spoken signals.

3. Experimental Design

3.1. The Human Judgement Task

To test whether different definitions of surprisal reflect perception of dialogue in the lexical channel of written and spoken conversations, we present a novel dialogue continuation acceptability judgement task. We present participants with a segment of a dialogue c, followed by a potential upcoming turn r. Following evidence that acceptability judgements are intrinsically gradient [16, 37], participants rate how plausible r is in the context of c on a scale of 1-5 ("Very Unlikely" – "Very Likely").

This task is similar to sentence acceptability judgements which have been widely studied in the context of surprisal. However, using dialogue turns as a base unit allows us to explore whether surprisal is predictive of acceptability perception in both written and spoken dialogues.

3.2. Data

Experiments on spoken dialogue were carried out using the Switchboard Telephone Corpus [38] which consists of over 2,400 chit-chat style conversations between 542 participants covering 70 topics. The corpus includes manual transcriptions and turn segmentations. These telephone conversations are an ideal data source for this task as speech is spontaneous and compared to other dialogue domains such as interviews, turns are relatively short, providing a diverse set of upcoming turns from which to sample. We carry out written-dialogue experiments on the DailyDialog corpus. This corpus includes 13,100 written conversations intended to resemble conversations from "daily life" [39] and thus provides a good match for Switchboard. Dialogues were extracted from web pages for English learners and, similar to Switchboard, span a broad range of topics.

3.3. Language model surprisal

We obtained surprisal estimates using the TurnGPT architecture [27], a variant of the GPT-2 [40]. Previous works which investigate the relationship between different language models and human comprehension behaviour consistently demonstrated that GPT-2 outperforms other comparable language model families [13, 15]. TurnGPT is trained with cross-entropy loss and uses an augmented input of three embeddings: token, position and speaker id, with the latter providing important cues for dialogue turn structure.

We took several steps to ensure comparability between spoken and written dialogue surprisal estimates. We used the GPT-2 BPE subword vocabulary, avoiding domain differences from out-of-vocabulary tokens (50259 tokens). We also removed punctuation except for turn-segmentation from DailyDialog. Our TurnGPT model was trained from scratch on equal amounts of data from DailyDialog and Switchboard ($\sim$ 4M tokens total). We used a slightly smaller architecture compared to the originally published model, with 8 layers, 8 attention heads, and an embedding size of 256. To verify that our model does not overfit, we checked that a 4 layer/4 head model didn't obtain lower perplexity on the validation set. The model was trained to achieve the lowest cross-entropy on a validation set containing equal proportions of data from Switchboard and DailyDialog.

The model achieves modality-specific perplexities of 60.27 68.31 on DailyDialog and Switchboard, respectively. Surprisal estimates are scaled by this modality ratio to adjust for inherent differences in predictability across corpora.

3.4. Behavioural study and Participants

To study a wide range of realistic instances of communication, stimuli were generated from each corpus by first obtaining contexts with a comparable quantity of information, operationalised as the cumulative per-token surprisal of a set of turns. This context surprisal measure was normalised by the corpus perplexity ratio (see Section 3.3) to enable compar-

Table 1: *Surprisal measure definitions: r and c are response and context word sequences resp.*

$$S_{total}(\mathbf{r}|\mathbf{c}) = \sum_{n=1}^{N}[S(r_n|\mathbf{r}_{<n}, \mathbf{c})]$$

$$S_{mean}(\mathbf{r}|\mathbf{c}) = \frac{1}{N}\sum_{n=1}^{N}[S(r_n|\mathbf{r}_{<n}, \mathbf{c})]$$

$$S_{relative}(\mathbf{r}|\mathbf{c}) = S_{mean}(\mathbf{r}|\mathbf{c}) - S_{mean}(\mathbf{r})$$

$$S_{max}(\mathbf{r}|\mathbf{c}) = \max[S(r_n|\mathbf{r}_{<n}, \mathbf{c})]$$

$$S_{var}(\mathbf{r}|\mathbf{c}) = \frac{1}{N-1}\sum_{n=2}^{N}[S(r_n|\mathbf{r}_{<n}, \mathbf{c}) - S(r_{n-1}|\mathbf{r}_{<n-1}, \mathbf{c})]^2$$

isons between the written and spoken stimuli. 10 contexts from each modality were sampled. Each was used to create 10 (*dialogue context c, upcoming turn r*) stimuli, 1 with the true upcoming turn, and 9 with negative upcoming turns. Negative turns were sampled to span the range of conditional surprisals expected from true (c, r) pairs. In total, 100 stimuli were generated for each modality [1].

52 participants were recruited from Prolific Academic, all were native English speakers based in North America. Each participant was presented with 25 stimuli through a Qualtrics survey, taking 9 ± 2 minutes to complete. Attention check questions that were manually selected as extremely likely (including noun overlap) and unlikely were interspersed throughout each survey. 20 participants obtained less than 75% overall accuracy on the check questions; their results were excluded.

3.5. Surprisal on a turn level

Token-level surprisal estimates were obtained from TurnGPT as the cross entropy loss between predicted and true tokens. From token-level surprisals, we compare a number of 'global' and 'local' operationalisations of turn surprisal suggested in previous work, summarized in Table 1 and described below.

Global metrics. Cumulative surprisal, $S_{total}(\mathbf{r}|\mathbf{c})$, is often used to model processing effort of an utterance [15]. To eliminate the influence of sentence length, we consider average surprisal per token, $S_{mean}(\mathbf{r}|\mathbf{c})$ [16]. We also consider the difference between the conditional and isolated mean utterance surprisal, $S_{relative}(\mathbf{r}|\mathbf{c})$, to control for the inherent surprisal of an utterance. Psycholinguistic theories diverge on whether or not discourse comprehension involves a context-independent analysis before integrating wider discourse context [41]. The isolated utterance surprisal, $S_{mean}(\mathbf{r})$, is computed as the average surprisal of response turn $\mathbf{r}$ conditioned on 100 randomly sampled contexts within the range of acceptable cumulative surprisal.

Local metrics. We are particularly interested in differences in information distribution between written and spoken language (cf [42, 7]) which may require the additional detail of local metrics. Thus, we consider maximum per-unit surprisal, $S_{max}(\mathbf{r}|\mathbf{c})$, as this has been used to capture points of extreme cognitive load [16]. We also quantify information distribution as surprisal variance between words, $S_{var}(\mathbf{r}|\mathbf{c})$ [15].

4. Results

Similar to previous works in sentence processing, we examine the relationship between these definitions of surprisal and judgement scores using ρ correlation [16, 15], as well as ordinal regression models.

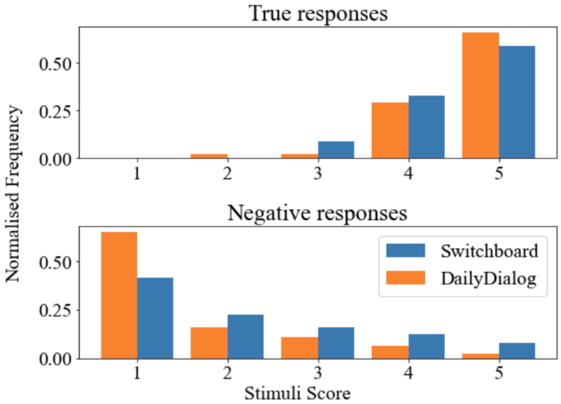

Figure 1: *Score distributions between corpora*

Table 2: *Correlation between surprisal and median judgement scores in DailyDialog and Switchboard*

Surprisal	**DailyDialog**		**Switchboard**	
	ρ	p-val	ρ	p-val
Total	-0.341	0.001	-0.273	0.006
Mean	-0.350	<0.001	-0.299	0.003
Relative	-0.360	<0.001	-0.262	0.009
Max	-0.407	<0.001	-0.400	<0.001
Variance	-0.295	0.003	-0.217	0.038

4.1. Perceptual task

Figure 1 demonstrates that participants were able to distinguish true turns from negatives samples; 95% and 90% of scores for true turns were either $[4, 5]$ for DailyDialog and Switchboard, respectively. Using the highest mean score per stimuli as a proxy for turn selection, participants obtained respective accuracies of 90% and 70%.

Figure 1 also highlights differences between the spoken and written corpora, particularly that participants make less certain judgements for Switchboard stimuli. Although score distribtions for true stimuli are similar across corpora, negative stimuli from Switchboard receive a wider range of scores – twice as many were rated as likely ($[3, 4]$) in Switchboard, and participants were more likely to rate turns as "Very Unlikely" ($[1]$) in DialyDialog stimuli. Because the informativeness of our stimuli context was controlled for, differences in score distributions are likely the result of turn characteristics. This suggests differences in the informative nature of basic turn units between the lexical content of spoken and written dialogue, which need to be explored in further.

4.2. Quantifying predictive power of surprisal

Given that people could leverage the stimuli context to accurately discriminate true upcoming turns from false ones, we used the plausibility scores to explore the relationship between human judgements and LM-estimated surprisal characteristics. If surprisal correlates with perception, we should expect responses that are surprising in context to obtain lower scores, i.e., a negative relationship between surprisal and score.

Results in Table 2 demonstrate weak but statistically significant negative correlation of median score with our operational-

[1] We provide examples of our stimuli: `https://sarenne.github.io/is-2022/`

Table 3: *Ablation from full model: Significant differences (> 2SE) are in bold.*

Removed feature	ELPD diff	SE
S_{total}	+1.9	**0.4**
$S_{relative}$	+0.7	0.9
S_{mean}	+0.3	1.2
S_{max}	-9.6	**4.3**
S_{var}	-2.1	1.9
Corpus	+0.3	1.2
All surprisal	-39.0	**10.5**

isations of surprisal. Slightly stronger correlation coefficients have been reported for related judgements of grammatical acceptability [15, 16], which likely reflect differences between tasks. The reasoning required to make predictions about dialogue (e.g., pragmatic/interaction features) may differ from factors involved in making isolated syntactic judgements.

Interestingly, previous works using surprisal estimates from similar language models find global operationalisations to offer more explanatory power than local ones [15]. Our results show the maximum conditional surprisal per token to have the strongest correlation with judgements scores. Again, this variation may be explained by differences in the reasoning required for these behavioural tasks.

Because the individual surprisal measures had significant but relatively weak correlations with median scores, we examine whether they could provide more information in combination. Since our perception experiment used a categorical rating scale (our predictee), we fit multilevel cumulative ordinal regression models (logit link function, uniformative flat priors), using the R package `brms` [43, 44]. Our predictors include the 5 surprisal metrics and a corpus indicator. We also include the context surprisal, and unigram and bigram overlap between the context and response (weighted by corpus frequency) to control for these potential sources of variation. Similarly, we include group level effects (i.e., random intercepts) to control for for participant and context identity. For brevity, we don't report group level effects here except to note that we consistently see non-zero variance associated with members of those groups. We evaluate models using leave-one-out cross-validation, estimating Expected Log Predictive Density (ELPD) [45].

We perform ablation of the individual surprisal measures with respect to the model using all predictors to investigate their contributions to model fit. Table 3 shows the difference in ELPD with respect to the full model (ELPD diff), as well as the associated standard error of the difference (SE). We take a model to have a significantly better fit when the ELPD difference is more than twice the SE. Removing all surprisal measures (leaving only n-gram and group predictors) significantly decreases the fit, indicating that participants were not making decisions based only on direct lexical matching between context and turn. In general, removing local measures (particularly S_{max}) reduces the fit, while removing global measures improves it, though not all differences were significant. Removing the corpus indicator also potentially reduces model fit, though the change is within the error margin.

Table 4 shows the difference between the full model, models with local or global surprisal metrics, and combinations of the two. The model including only global features performs worse than the full model, suggesting that their inclusion (specifically S_{total}) is leading to overfitting. However, the difference is not significant. Our best-fitting model includes local (S_{max}, S_{var}) metrics with S_{mean}. This indicates that a combination of individually weak predictors needs to be considered in determining the perceived likelihood of a turn in a specific context. The 95% confidence intervals for surprisal measure coefficient estimates and the corpus indicator all exclude zero (while context surprisal and n-gram measures confidence intervals include zero), supporting a non-zero contribution for these effects. The estimates indicate that stimuli with higher turn surprisals (S_{mean}, S_{max}) received lower ratings. However, contrary to the correlation analysis, higher word-to-word surprisal variability (S_{var}) contributes to higher scores when accounting for the other metrics. We did not find any further improvement in model fit from including corpus interaction terms with the surprisal measures.

Table 4: *Differences between models with full, local, and global features. *corpus adds an interaction between the corpus indicator and the surprisal measures for that model.*

Included Features	ELPD diff	SE
global	-8.4	4.4
local	+0.6	2.2
local+relative	+1.3	1.2
local+mean	**+2.2**	**1.0**
*(local+mean)*corpus*	+0.4	2.1

5. Discussion and Conclusions

These results have demonstrated that people can make accurate judgements about upcoming utterances in both written dialogue and transcripts of spoken dialog based on a fixed amount of context, but do so less effectively in the later. They confirm the utility of our task for studying perception of dialogue and indicate that lexical information is distributed differently between written and spoken dialogue (potentially an effect of the different channels available in these modalities). The results also probe the perceptual validity of the response selection paradigm used throughout conversational language modelling [26, 29]. Although people can accurately discriminate between true and false upcoming turns, there is often more than one plausible response for a given context [46].

We then explored the predictive power of global and local operationalisations of surprisal for this communication-based task. In some ways, our findings are complementary to previous works – all operationalisations displayed weak but significant correlation with human judgements across written and spoken dialogue. However, we found differences in the respective utility of local and global surprisal compared to previously reported results on monologue-based tasks [16, 15, 13].

Combinations of global and local operationalisations provided the highest predictive power for our task. Surprisingly, once other sources of variation had been accounted for, S_{var} had a positive effect on the fit of the logistic regression model such that more variance between turn tokens produces higher rating. Further investigation of the utility of different surprisal metrics and their interactions is required.

Stimuli used in this work were sampled within a fixed range of cumulative context suprisal. People have been shown to make effective predictions in dialogue using very little context [47], but testing different amounts of context could provide insight into how much information people use to make judgements. Given that LM architecture is known to affect the explanatory power of surprisal estimates, future work could also compare estimates from different architectures [13, 15, 16].

6. References

[1] M. Schrimpf, I. A. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. G. Kanwisher, J. B. Tenenbaum, and E. Fedorenko, "The neural architecture of language: Integrative modeling converges on predictive processing," *Proceedings of PNAS*, vol. 118, 2021.

[2] M. H. Christiansen and N. Chater, "The now-or-never bottleneck: A fundamental constraint on language," *Behavioral and brain sciences*, vol. 39, 2016.

[3] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences*, vol. 27, pp. 169 – 190, 2004.

[4] B. V. Tucker and M. Ernestus, "Chapter 4. Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon," in *Polylogues on The Mental Lexicon*, 2021, pp. 77–108.

[5] K. Ethayarajh, "How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings," in *EMNLP*, 2019.

[6] Q. Liu, M. J. Kusner, and P. Blunsom, "A survey on contextual embeddings," *ArXiv*, vol. abs/2003.07278, 2020.

[7] M. P. Aylett and A. Turk, "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech," *Language and Speech*, vol. 47, pp. 31 – 56, 2004.

[8] J. Hale, "A probabilistic earley parser as a psycholinguistic model," in *NAACL*, 2001.

[9] R. Levy, "Expectation-based syntactic comprehension," *Cognition*, vol. 106, no. 3, pp. 1126–1177, 2008.

[10] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[11] S. L. Frank, L. J. Otten, G. Galli, and G. Vigliocco, "The ERP response to the amount of information conveyed by words in sentences," *Brain and language*, vol. 140, pp. 1–11, 2015.

[12] R. Levy, "Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results," in *ACL*, 2011, pp. 1055–1065.

[13] E. G. Wilcox, J. Gauthier, J. Hu, P. Qian, and R. P. Levy, "On the predictive power of neural language models for human real-time comprehension behavior," *ArXiv*, vol. abs/2006.01912, 2020.

[14] A. Goodkind and K. Bicknell, "Predictive power of word surprisal for reading times is a linear function of language model quality," in *Proceedings of CMCL*, 2018, pp. 10–18.

[15] C. Meister, T. Pimentel, P. Haller, L. Jäger, R. Cotterell, and R. Levy, "Revisiting the uniform information density hypothesis," in *Proceedings of EMNLP*, 2021.

[16] J. H. Lau, A. Clark, and S. Lappin, "Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge," *Cognitive Science*, vol. 41, no. 5, pp. 1202–1241, 2017.

[17] S. Richter and R. Chaves, "Investigating the role of verb frequency in factive and manner-of-speaking islands," in *CogSci*, 2020.

[18] A. Warstadt, A. Singh, and S. R. Bowman, "Neural Network Acceptability Judgments," *Transactions of ACL*, 2019.

[19] T. Kuribayashi, Y. Oseki, T. Ito, R. Yoshida, M. Asahara, and K. Inui, "Lower perplexity is not always human-like," in *ACL*, 2021.

[20] A. Goodkind and K. Bicknell, "Local word statistics affect reading times independently of surprisal," *ArXiv*, vol. abs/2103.04469, 2021.

[21] R. Futrell, E. Gibson, and R. P. Levy, "Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing," *Cognitive Science*, vol. 44, 2020.

[22] G. Kharkwal and S. Muresan, "Surprisal as a predictor of essay quality," in *BEA@ACL*, 2014.

[23] H. H. Clark and D. Wilkes-Gibbs, "Referring as a collaborative process," *Cognition*, vol. 22, pp. 1–39, 1986.

[24] V. M. Silva and M. Franke, "Pragmatic prediction in the processing of referring expressions containing scalar quantifiers," *Frontiers in Psychology*, vol. 12, 2021.

[25] J. Degen and M. K. Tanenhaus, "Processing scalar implicature: A constraint-based approach," *Cognitive science*, vol. 39 4, pp. 667–710, 2015.

[26] M. Henderson, I. Casanueva, N. Mrkvsi'c, P. hao Su, Tsung-Hsien, and I. Vulic, "ConveRT: Efficient and accurate conversational representations from transformers," *ArXiv*, vol. abs/1911.03688, 2020.

[27] E. Ekstedt and G. Skantze, "TurnGPT: a Transformer-based Language Model for Predicting Turn-taking in Spoken Dialog," in *Findings of ACL: EMNLP*, 2020, pp. 2981–2990.

[28] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, "Transfertransfo: A transfer learning approach for neural network based conversational agents," *ArXiv*, vol. abs/1901.08149, 2019.

[29] J. Han, T. Hong, B. Kim, Y. Ko, and J. Seo, "Fine-grained post-training for improving retrieval-based dialogue systems," in *NAACL*, 2021.

[30] C. Liu, R. Wang, J. Liu, J. Sun, F. Huang, and L. Si, "DialogueCSE: Dialogue-based contrastive learning of sentence embeddings," in *Proceedings of EMNLP*, 2021, pp. 2396–2406.

[31] N. G. Ward and B. H. Walker, "Estimating the potential of signal and interlocutor-track information for language modeling," *Proceedings of Interspeech*, 2009.

[32] S. K. Kim and M. Sumner, "Beyond lexical meaning: The effect of emotional prosody on spoken word recognition," *The Journal of the ASA*, vol. 142, no. 1, pp. EL49–EL55, 2017.

[33] B. M. Ben-David, N. Multani, V. Shakuf, F. Rudzicz, and P. H. van Lieshout, "Prosody and semantics are separate but not separable channels in the perception of emotional speech: Test for rating of emotions in speech," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 1, pp. 72–89, 2016.

[34] J. Hirschberg and J. Pierrehumbert, "The intonational structuring of discourse," in *ACL*, 1986, pp. 136–144.

[35] H. H. Clark and S. E. Brennan, "Grounding in communication." in *Perspectives on socially shared cognition.*, 2004, pp. 127–149.

[36] M. J. Sjerps, C. Decuyper, and A. S. Meyer, "Initiation of utterance planning in response to pre-recorded and "live" utterances," *Quarterly Journal of Experimental Psychology*, vol. 73, pp. 357 – 374, 2019.

[37] N. Chater, J. B. Tenenbaum, and A. L. Yuille, "Probabilistic models of cognition: Conceptual foundations," *Trends in Cognitive Sciences*, vol. 10, pp. 287–291, 2006.

[38] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of ICASSP 1992*, 1992, pp. 517–520.

[39] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of ICJNLP*, 2017, pp. 986–995.

[40] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, 2019.

[41] M. S. Nieuwland and J. J. A. V. Berkum, "When peanuts fall in love: N400 evidence for the power of discourse," *Journal of Cognitive Neuroscience*, vol. 18, pp. 1098–1111, 2006.

[42] R. Levy and T. F. Jaeger, "Speakers optimize information density through syntactic reduction," in *NIPS*, 2006.

[43] P.-C. Bürkner and M. Vuorre, "Ordinal regression models in psychology: A tutorial," *Advances in Methods and Practices in Psychological Science*, vol. 2, no. 1, pp. 77–101, 2019.

[44] P.-C. Bürkner, "brms: An R package for Bayesian multilevel models using Stan," *Journal of Statistical Software*, vol. 80, no. 1, pp. 1–28, 2017.

[45] A. Vehtari, A. Gelman, and J. Gabry, "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC," *Statistics and computing*, vol. 27, no. 5, pp. 1413–1432, 2017.

[46] A. Cervone and G. Riccardi, "Is this dialogue coherent? learning from dialogue acts and entities," in *SIGDIAL*, 2020.

[47] S. Wallbridge, P. Bell, and C. Lai, "It's not what you said, it's how you said it: discriminative perception of speech as a multichannel communication system," *Proceedings of Interspeech*, 2021.