Shared knowledge in natural conversations: can entropy metrics shed light on information transfers?

Eliot Maës, Philippe Blache, Leonor Becerra-Bonache

Shared knowledge in natural conversations: can entropy metrics shed light on information transfers?

Eliot Maës[1,3] **Philippe Blache**[2,3] **Leonor Becerra-Bonache**[1,3]
[1]Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
[2]LPL, CNRS & Aix-Marseille University
[3]ILCB (Institute of Language, Communication and the Brain)
{eliot.maes,leonor.becerra}@lis-lab.fr,
blache@ilcb.fr

Abstract

The mechanisms underlying human communication have been under investigation for decades, but the answer to how understanding between locutors emerges remains incomplete. Interaction theories suggest the development of a structural alignment between the speakers, allowing for the construction of a shared knowledge base (*common ground*). In this paper, we propose to apply metrics derived from information theory to quantify the amount of information exchanged between participants, the dynamics of information exchanges, to provide an objective way to measure the common ground instantiation. We focus on a corpus of free conversations augmented with prosodic segmentation and an expert annotation of thematic episodes. We show that during free conversations, the amount of information remains globally constant at the scale of the conversation, but varies depending on the thematic structuring, underlining the role of the speaker introducing the theme. We propose an original methodology applied to uncontrolled material.

1 Introduction

Theories of interaction explain how participants elaborate their discourse in the perspective of exchanging information, executing a task, establishing a joint action, etc. These theories stipulate in particular that such activity is correlated with the building of a shared knowledge between participant, also called *common ground* (Pickering and Garrod, 2004, 2021). In these frameworks, the quality of an interaction depends on the capacity of building such mutual knowledge, which to its turn depends on the alignment of linguistic representations between participants. These mechanisms are based on different levels of convergence between the participants, that can occur at any level: lexical, syntactic, prosodic, as well as gestures, behaviors, etc. One hypothesis is that this phenomenon is also visible at the semantic level, showing a coordination between participants in terms of information exchange that can be uncovered by studying the amount of such information and its dynamics during a conversation.

The goal of this work is therefore to evaluate these questions by means of information-theoretical measures (Shannon, 1948): sharing information relies on the use of simple symbols which can be combined, concatenated to transfer increasingly complex knowledge. Moreover, it is possible to analyze the dynamics of this process, whether the amount of transfer vary during a conversation, at what position, and whether an alignment between participants can also be observed at this level. An estimation of the quantity of information exchanged between participants and its dynamics could therefore constitute an objective way to measure the common ground instantiation.

Several works have been done in this direction, based on lexical information measured by entropy, and showing a convergence between participants. Inspired by Xu and Reitter (2016), we study the dynamics of information transfer at three levels: first globally, at the scale of an entire conversation, by taking into account productions from both speakers into a same system. Doing that, we propose to identify whether some specific phenomena (e.g. peaks) appear in the amount of exchanged information and that could be related with discourse-level structures (e.g. topic shift). Second, we will study the global evolution of entropy for each speaker, trying to exhibit some convergence patterns (e.g. phase synchronization). Third, we propose to apply the same type of analysis at the scale of a topic, by studying the dynamics of information exchange within a topic (e.g. decrease of entropy) as well as the complementary patterns between speakers. Last, but not least, this is the first work in this domain applied to unrestricted natural conversations.

This paper presents several contributions, corresponding to important differences with the litera-

Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL), pages 213 - 227

ture. First, we propose to explore this question applied to free conversations instead of task-oriented or controlled ones. Second, in difference with existing works, we evaluate the dynamics of the exchange based on well-defined inter-pausal units instead of sentences (a not adequate notion for spoken languages). Finally, we base our analysis on thematic annotation made by an expert (human linguist) instead of an automatic topic segmentation.

The paper is organised as follows. In Section 2, we review the different approaches to these questions in the litterature. In Section 3, we describe our conversational dataset and the methodology we apply. Our experiments and a discussion of the results are presented in Section 4.

2 Related Works

Several studies have proposed to use information-theoretic measures to study language processing. The general idea is to approach an evaluation of the cognitive load through quantitative estimation. In a seminal work, Hale (2001) introduced the notion of *surprisal*, defined as the negative log-probability of a word given the preceding context, to measure processing difficulty. This approach has been picked up by many studies in psycholinguistic, showing in particular a correlation between reading times and surprisal (Monsalve et al., 2012; Frank et al., 2015). In the same vein, based on grammatical probability distributions, entropy reduction has been proposed to evaluate the informational contributions of each word as a complexity processing measure (Hale, 2016). At the lexical level, without any additional syntactic information than what is understood by the linguistic model, entropy has been proposed to estimate sentence information content in discourse. We offer in this section an overview of the main works done in this direction by first presenting the main approaches to measure information content and second the methods for studying variations of such measures at the discourse level.

2.1 Measuring Information Content

In discourse, each lexical choice can be described as a random variable X_i that is constrained by a number of influences, both short range (sentence structure, local topic) and long range (global context). As the relevant context builds up, the next word prediction is assumed to become easier and easier as more contextual cues are available to the discussion. The information density of this random variable is estimated as the entropy $H(X_i)$ defined by Shannon (1948). We especially follow Xu and Reitter 2018; Giulianelli et al. 2021 in modeling the information content.

The influence of the local context on the word choice is typically modelled at utterance or sentence level with conditional probabilities; sentence entropy is taken as the average entropy of the words comprising that sentence. Therefore, for a given sentence S comprising of a sequence of n words $w_1, w_2, ...w_n$

$$H(w_1...w_n) = -\frac{1}{n}\sum_{w_i \in S} logP(w_i|w_1...w_{i-1}) \tag{1}$$

Keller (2004) and Genzel and Charniak (2003) exposing a correlation between sentence length and entropy values, we also compute a *normalized* version of our entropy metric to remove dependence to sentence length, by dividing the previously computed metric by the average obtained on all sentences of the same length:

$$H'(S) = \frac{H(S)}{\frac{\sum_{W \in L(n)} H(W)}{\#\{L(n)\}}} \tag{2}$$

where $L(n)$ is the set of sentences of length n, ie sentences of the same length as our sentence S.

The initial studies use n-gram language models to estimate word probabilities, which fail to take more long range dependencies into account. The natural reaction is to question the effect of context, which is the approach taken by Giulianelli et al. (2021). They introduce the distinction between *decontextualised* entropy, that only uses the local sentence S as context, and *contextualised* entropy, which utilises the global context C, i.e. all previously mentioned sentences up to the current word, as context. The contextualised entropy of a word is therefore computed as the conditional entropy of a word depending on both the local and global context.

The difference between the amounts of information at the local and global contexts if carried by the mutual information term $MI(S|C)$:

$$H(S|C) = H(S) - MI(S|C) \tag{3}$$

2.2 Entropy variations in language processing

Genzel and Charniak (2002) proposed the *entropy rate constancy* principle stipulating that the rate of transmitted information remains approximately

constant. Initially enunciated for written texts, this principle has been applied to natural conversation, albeit with some adaptations.

Following the entropy rate constancy principle, the conditional entropy remains constant through the dialogue. As a consequence, local entropy and mutual information have to vary in the same proportions. At the scale of a dialogue, it has been shown that the two arguments of this equation regularly increase in the same way (Genzel and Charniak, 2002). But at the same time, even though the entropy should remain constant throughout the dialogue, local variations are possible. This aspect has been explored by studying the entropy at specific positions, taking into account the role of the participants in the conversation (Xu and Reitter, 2016, 2018; Giulianelli and Fernández, 2021). These studies are based on a segmentation of the discourse in a sequence of separate topics, with the idea that this succession of thematic episodes could be associated with a variation in the entropy. In this perspective, Qian and Jaeger (2011) has shown a correlation between entropy decrease and potential topic shift in written text: topic shift corresponds to the drop of the mutual information term. More recently, Xu and Reitter (2016) exhibited a symmetry in the entropy fluctuations within a topic depending on the speakers' roles. A new topic corresponds to introducing new information into the context, which means high entropy at the beginning of a topic for the speaker who introduces it (topic *initiator*). Reciprocally, their partner (called in these studies *responders*) progressively update the context, which means that for them, entropy starts low and progressively increases until the next topic. As a consequence, these fluctuations show a convergence pattern between interlocutors within a topic.

3 Datasets and Models

3.1 Datasets

Previous work on information density focusing mostly on task-related conversational datasets such as MapTask (Anderson et al., 1991), we explore whether conclusions drawn on such specific data further generalise to natural conversation by applying the same methods on the Paco-Cheese corpus (Priego-Valverde et al., 2020; Amoyal et al., 2020). Indeed, since vocabulary is not as controlled in natural conversations as it is in tasks, the conversation might drift onto less predictable topics that rely more on common knowledge.

Paco-Cheese (PC) (Priego-Valverde et al., 2020) is a multimodal corpus containing audio and video recordings of 26 interactions between dyads of participants. Conversations are in French and lasting 15 to 20 minutes. Participants were given a short prompt to read to elicit conversation but were otherwise free to talk about the topics of their choice. About half (16) of the conversations happened between participants that were not acquainted. Manual transcription was obtained, then automatically aligned to the audio signal and segmented using SPPAS (Bigi, 2012). Consequently, the speech segments we consider here are inter-pausal units (IPUs) - segments boundaries are defined by pauses longer than 200ms of silence - which commonly are shorter than sentences. The corpus is also enriched with annotations for noise, laugh, pauses, feedbacks, head nods and smiles (Amoyal, 2018; Amoyal and Priego-Valverde, 2019). Expert thematic annotation has been added to 16 of the dialogues. Excerpts from the corpus can be found in Appendix A.

Relying on these annotations, we compute information content values for the dialogues and consider its evolution at two levels: *global evolution* throughout the conversation, and *local evolution* in a given conversational theme.

3.2 Language Models

We estimate information content throughout the dialogue by computing per-word entropy for each sentence, using language models trained on different corpora and finetuned on the dataset of interest.

Previous works relied both on n-gram models (Xu and Reitter, 2018) and Transformer models (Giulianelli et al., 2021). Models were then not straightly compared however the latter method provides with two advantages: first, Transformers allow for the possibility to take larger amounts of contextual information into account; second, default Tokenizers in the pipeline are trained using a Byte-Pair Encoding, which allows them to properly deal with out-of-vocabulary (OOV) tokens. Those rarer words would be especially important in predicting surprise and information content in the conversation.

After experimenting with n-gram models, RNNs and the GPT-2 language model (Radford et al., 2019) - we disregard more recent models using masking-based learning in order to focus on more

Table 1: Perplexity for the models used compared to that of GPT-2 pretrained models

model	lang.	pretraining	finetuning	perplexity	OOV
SRILM	FR	decoda	x	132,32	0.5%
RNN	FR	wikipedia	x	83,16	-
GPT-2	FR	wikipedia		125,39	-
GPT-2	FR	wikipedia	x	32,51	-

cognitive-plausible models - we chose to focus on the latter as GPT-2 demonstrates both lower perplexity and has been shown to better correlate with human surprisal in language understanding (Michaelov et al., 2021). We rely on HuggingFace's implementation of the model[1], using default tokenizers and parameters (Wolf et al., 2020). Finetuning is required to adapt the language model from written input to the specificities of natural conversation. We therefore finetune the models on a 70% split of each target corpus. As shown in Table 1, finetuning yields a substantial reduction in the model's perplexity.

The information content of an utterance is computed sequentially, using log-probabilities predicted by the model for each token in the sentence. Several lengths of context are considered (current utterance only $H(S)$; several utterances; every preceding utterance $H(S|C)$) and Mutual Information is computed from the difference between $H(S)$ and $H(S|C)$.

More information on models parameters and finetuning can be found in Appendix B.

3.3 Statistical Models

With our experiments, we study the dynamics of information transfers at two levels: i) globally, at the level of the entire conversation; ii) locally, at the level of topic episodes. We fit linear models on information content estimated by the language models on those two conditions. In those models, the logarithm of the information content is the response variable (*H(S|C)* or *H(S)*) and the logarithm of turn position (whether global, `logp`, or relative to the local theme, `logt`) is the fixed effect. Dialogues are considered a random effect in this analysis.

We also include in our analysis a comparison between utterance lengths to validate that using IPUs does not affect the conclusions we draw from the data.

[1] `https://huggingface.co/gpt2`, using weights from `dbddv01/gpt2-french-small` for the french model

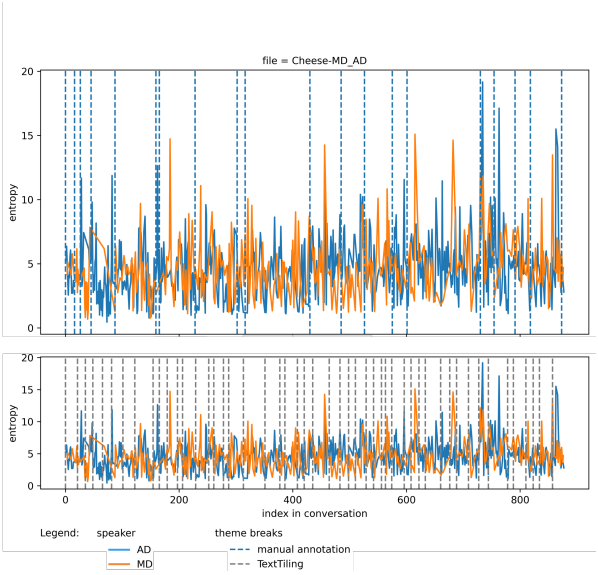

Figure 1: Evolution of normalised contextualised entropy on one example dialogue. The two speakers are plotted in different colors. Dashed lines indicate the start of new themes in the manual annotation (top) and automated annotation (bottom).

3.4 Peak identification and correlation to thematic annotation

Topic Segmentation Information content evolution is typically studied at the dialogue level (global context), but also locally, at the level of *topic episodes*. Annotations for this partitioning can be derived automatically using tools such as TextTiling (Hearst, 1997). This algorithm relies on lexical co-occurrences patterns to compute a similarity score between sentences and segment a text into subtopic shifts.

To complement the manual annotation of themes in Paco-Cheese, we obtain automatic extraction of theme changes using NLTK's implementation[2] of the TextTiling algorithm. This step furthermore allows to compare human sensitivity to topic change to lexical changes (see Figure 1), an analysis which has not been done on the corpus yet.

Entropy Peak Detection and Analysis Investigating the location of information exchanges, we consider peaks of entropy as potential locations for the introduction of new data to the conversation. Assimilating those values to outliers, two unsupervised methods are used to detect those values. Entropy series are detrended and scaled before

[2] `https://www.nltk.org/api/nltk.tokenize.html#module-nltk.tokenize.texttiling`

Table 2: Estimates and significance for the effect of position on information content for the linear mixed-effect models on Paco-Cheese

		interaction		theme	
global	log position	0.027	***		
	log position			-0.032	***
initiator	log position			-0.021	**
follower	log position			-0.015	

further computations. The first method of outlier detection involves local detection of unusual values; we rely on `scikit-learn` (Pedregosa et al., 2011) implementation of Local Outlier Factor for this. The second method (hereafter NormOutlier) involves globally comparing the values and selecting the highest two percent. For both methods, parameters were chosen as optimal values on a subset of the data based on accuracy, precision and recall metrics. We finally compare the performances of those methods in predicting thematic episodes boundaries, to basic classifiers from the `scikit-learn` dummy module.

We also leverage Part-of-Speech tagging and Feedback annotations from the dataset to explore which words are most unexpected for the model.

4 Experiments

In this section we present the results of our experiments with the Paco-Cheese dataset. Taking the values of $H(S)$ (i.e., the information content of a sentence) and $H(S|C)$ (i.e., the contextualised entropy) estimated by the language model, we also compute the difference between contextualised and decontextualised entropy (MI). We extend results obtained by previous works with this new corpora containing free conversations. We then explore those results using qualitative and quantitative analysis of locations with high information content.

4.1 Speakers behavior in natural conversation

Global evolution We find a positive effect of turn position on information content when taking the entire Paco-Cheese dialogues as the context unit (see Table 2). This effect can however be entirely attributed to the structure of the corpus as conversation usually start with a few sentences of explanation of the experiment and two one-sided readings of the jokes. Indeed when focusing only on the free conversation, we find that this positive effect disappears (see Figure 2 for the difference of entropy evolution between the two conditions).

Local evolution: themes We do however observe an effect of turn position on information content at the level of themes (β = -0.032, $p < 0.001$) (see Figure 3), which seems to be entirely driven by the behavior of the topic initiator (β = -0.021, $p < 0.001$). We observe no effect of turn position on information content for the other locutor responding to the topic initiation.

We attribute the lack of overall effect of position to the structure of the conversation, as in a natural paradigm speakers will naturally shift from one topic to the next, without necessarily relying on previously mentioned context to move the conversation forward. Themes, however, make up smaller, coherent units of a conversation. The negative effect of turn position on information content in themes would seem to be going against the principle of Uniform Information Density (Jaeger and Levy, 2006) and its applications to dialogue which indicate that information content should be increasing; it is however in line with Xu and Reitter (2018)'s findings that the information content will be either constant or slightly decreasing the more the topic progresses. We postulate that the reason why we do not observe an effect of position is because the responder is active in helping constructing the theme and does not simply fall back into a passive role at the introduction of a new topic.

The full results of the statistical analysis and accuracy of theme change detection can be found in Appendix C.[3]

4.2 Units of sense in a conversation: IPUs vs. sentences?

Unlike other works that compute entropy at the level of a "sentence" (which is not valid when studying spoken language), the input to our models are inter-pausal units (speech separated by 200ms pauses). IPUs being shorter than sentences or turns and potentially made of fewer words, they offer the possibility of a finer granularity, more in line with linguistic characteristics of dialogues.

One might expect this change of scale to affect the patterns displayed in information content, as longer interventions would bring in more information at once. Differences between topic initiator and responder might appear more strongly with a more frequent use of short utterances and feedback.

[3]Codes and statistical analysis are available at `https://github.com/ejmaes/multimodal-itmodels`

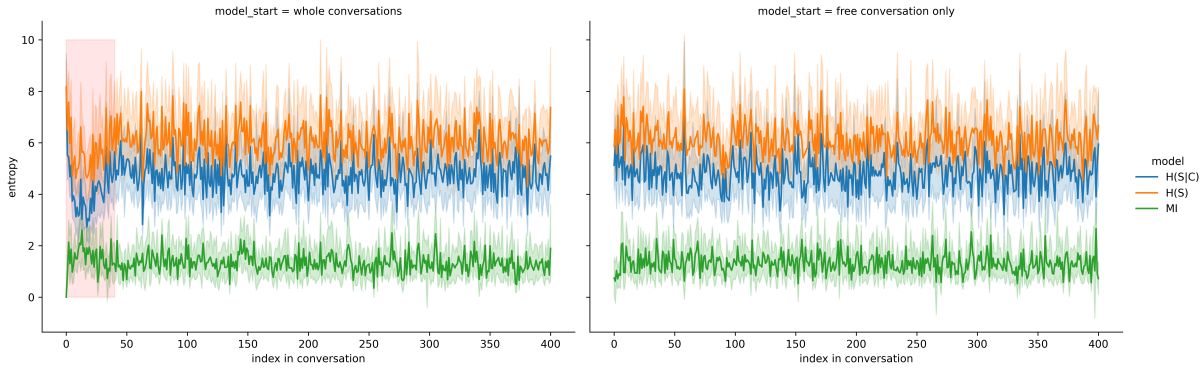

Figure 2: Average evolution of the entropy throughout the conversation for the Paco-Cheese corpus. Left: starting at file start; Right: removing introductions and prompt reading to start analysis at the beginning of the free conversation. In red, the approximate duration of conversation starters (varies between dialogues)

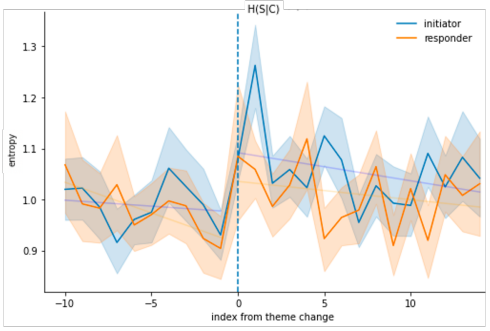

Figure 3: Average information content in the utterances surrounding the start of a new theme, for both speakers.

To test this hypothesis, we aggregated IPUs by a given speaker that were separated by silences shorter than 1 second and were not interupted by the other speaker. The obtained utterances are akin to sentences in terms of length and semantic content. For the comparison to be more accurate, we remove the IPUs of the first part of the dialogs, which correspond to the reading of jokes and not to actual conversation. We then fed this new data into the language model. Results (see Table 3) mostly appear robust to the aggregation, with a main effect of position on entropy at the level of themes and for the speaker initiating the theme.

Table 3: Comparison between IPUs and sentences - Estimates and significance for the effect of position on information content

		IPUs		sentence - 1s	
global	both speakers	0.015		-0.22	**
theme	both speakers	**-0.030**	***	**-0.029**	***
	initiator	**-0.024**	**	**-0.033**	**
	follower	-0.014		-0.017	

4.3 Distribution of entropy peaks against themes

The distribution of information in the conversation, despite being stable on a global level, is not smooth on a local scale, as the even flow of entropy is sometimes intersected with peaks of local uncertainty. We ponder whether those peaks only correlate to endemic features of the conversation, such as the introduction of new information to the discussion, or whether they inform on model shortcomings that need to be addressed to better understand the characteristics of information transmission and common ground instantiation in conversation.

4.3.1 Theme change in conversation: smooth or abrupt behavior?

Inspired by the behavior observed in entropy values around theme breaks (see Figure 3) and the decrease in entropy for the initiator throughout the theme they introduced, we wonder whether it is possible to predict theme breaks from entropy values and more specifically entropy peaks.

We first start by exploring how similar automatic and manual annotations actually are. A first quantitative approach reveals that TextTiling systematically overestimates the number of themes by conversation in our dataset (Figure 4), predicting 565 thematic episodes whereas the dataset only has 268 (see Table 4). This might be an indicator of the existence of subtopics in the conversations; however, locations indicated by TextTiling as the start of new themes only weakly correlate with expertly annotated locations. A first hypothesis as to explain those results involves the existence of *transitions* phases in-between two thematic episodes. Transitions are frequently annotated in the corpus, with

Table 4: Average number of themes per dialogue in each dataset, as annotated vs estimated by TextTiling

	Annotations	TextTiling
PACO-CHEESE	16.4 ($\pm$ 2.8)	34.5 ($\pm$ 7.0)

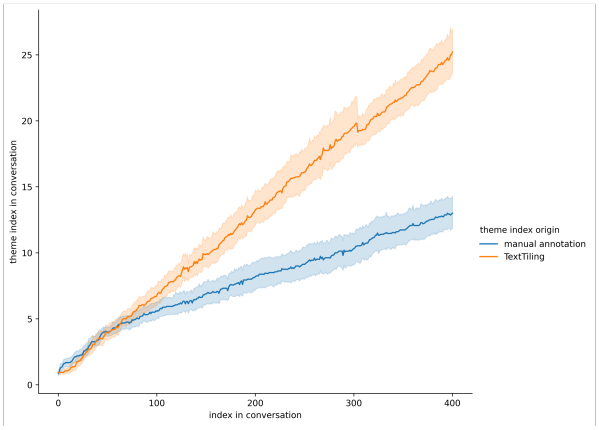

Figure 4: Average number of themes discussed in the conversation as a function of IPU index, according to manual annotation (blue) and automatic annotation (orange). Some conversations are shorter than others, which might cause the average number of themes to drop around some indexes.

13.6 $\pm$ 3.3 transitions per conversation, slightly less than the number of themes per dialogue. Indeed, if transitions are annotated, then boundaries between themes must be considered to be flexible enough. We then consider that a prediction falling within a small window of a boundary will be an accurate one; this yields better results, despite prediction accuracy remaining quite low (80 common locations out of 288 annotated theme changes, see Table 5).

Manual and automatic annotation therefore appear to consider different features and rules to establish thematic boundaries. But if automatic annotation is more sensitive to new vocabulary introduced in the conversation to label thematic changes, we hypothesize might also correlate more with entropy values. For this reason we compare the location of information content peaks to the distribution of topics both manually and automatically predicted.

Peak location does not accurately predict theme changes or TextTiling results, though correlating peaks to TextTiling yields slightly better results than manual annotation. For manual annotation, Local Outlier detection allows for the detection of the largest number the theme changes ($precision = 0.172$ / $recall = 0.65$ within a 5 IPUs window), predicting a larger number of locations of interest than annotated. Peak detection further correlates with automated annotation of theme changes, which further support the hypothesis of entropy peaks appearing around locations where new content is introduced. This method for predicting thematic boundaries however does not fare better than a baseline classifier trained directly on entropy values and sentence length to detect topic boundaries.

4.3.2 Language models and natural conversation

To further analyze model and participants behavior throughout the conversation, we shift our focus to per-word entropy. We focus on two aspects: words with high entropy on the one hand, and the way the model deals with conversational feedback on the other.

From peak locations, a set of vocabulary with the highest entropy values is extracted. We cross this list with part-of-speech tagging and feedback annotation available in the corpus before going further. We note that most of those words are nouns, with the stronger occurrences being proper nouns, which is expected since those words wouldn't be known to the model - or, in the case of locations, logical in the conversation - prior to encountering them. Some of those unexpected words would however not be evaluated by the speakers as this significant, since they are already part as their shared knowledge (nearby locations, daily life abbreviations, names of known individuals...). Thus most of these words may simply be unexpected in this context or too unusual for the model, and do not provide any new information to the topic at hand. However, a small percentage of words do; and in the case of words reappearing later in the conversation, a slight decrease in entropy is observed. A list of unusual words with high entropy causing the appearance of peaks is provided in Appendix D.1.

We finally turn our attention to backchannels, a discourse-specific occurrence through which a listener can interact with the speaker and notify them on their thought process without requiring taking the floor. Backchannels typically include movements (head nods, smiles or facial expressions), small words (*yes, okay, no, sure*...) or short utterances that do not disrupt the conversation flow. A qualitative analysis of peak locations had revealed the presence of feedbacks among the utterances of interest; further inspection actually reveals this is not an issue in modeling. Indeed, most feedbacks generate lower than average entropy. But

Table 5: Comparison of manually annotated theme changes locations to peak locations and theme breaks according to automated annotation. A baseline classifier (DumStrat) trained to predict theme breaks is added for reference.

	features	target	best result	True Positive	precision	recall
TextTiling	text	manual annotation	window=5	80	0.136	0.299
LocalOutlier	entropy		window=5	173	0.172	0.646
NormOutlier	entropy		window=5	23	0.174	0.086
DumStrat	entropy + text features		-	261	0.268	0.113
LocalOutlier	entropy	TextTiling	window=5	381	0.278	0.674

sometimes longer feedbacks conveying meaning a bit more specific generate uncertainty, same as other utterances in the dialogue, with the difference than those productions from the listener are more concise than utterances from the speaker. It is especially the fact for unexpected, negative input, but makes perfect sense on a cognitive standpoint.

A more detailed view into feedbacks types, frequencies and related entropy is available in Appendix D.2.

5 Conclusion

The results presented in this paper represent a new contribution for the study of information exchange during conversation. First, this work only relies on free natural conversations, without adding more controlled corpora. In particular, in difference with other works in the literature, we do not add any task-oriented dialogue (such as MapTask) nor telephone conversation (such as Switchboard), that have known specific impacts on turn taking and topic shift. In terms of methodology, we decided to use a prosodic segmentation of the input (pauses longer than 200ms) generating identify inter-pausal units usually used in studies on spoken language. IPUs are discourse segments with a certain coherence only identified on the basis of the acoustic signal. These segments offer a finer-grained view of the input in comparison with the segmentation into sentences that are usually used in the literature. This notion of sentence is not only problematic when applied to spoken language (the existing works do not precisely explain to what they correspond), but may also introduce a bias when studying topic shift, these two segments being possibly the same. Finally, we are using with this analysis a thematic segmentation that was done manually by experts, rather than relying on automatic segmentation as previous works might have done. TextTiling identifies topics based on semantic similarity; here annotations are based on higher-level information, bringing together all different linguistic and non-verbal information, providing a much more reliable segmentation.

Our results first confirm that at the scale of a conversation, entropy remains stable, as it has been observed in other works. At a local level, when segmenting the discourse in themes, we also observe a specific behavior, showing a decrease in the entropy of the speaker introducing the theme, which is expected. However, no significant pattern can be observed for the responder, for who the entropy remains approximately stable. To be more precise, we did not observe any increase in the entropy. As a consequence, we cannot say that a convergence in the entropy rate between the different speakers can be observed at the scale of a theme. This result is important in the study of conversational interactions. It means that convergence between speakers, which is necessary during a conversation, is a complex phenomenon that cannot be observed only on the basis of quantity measures. At the same time, the analysis of entropy constitutes a robust cue for evaluating how much and when information is transferred between speakers in a natural setup; however it must be complemented with data from other sources to assist the model in isolating truly important sections of the dialogue, from noise (rarer words that are logical in the context).

This work opens the door to further study. For starters, as previously mentioned, enriching the models with information, coming from other modalities would most likely refine the analysis. Among the modalities of interest are audio (speech rate is known to be modulated according to the difficulty of the information), video (gaze), and cerebral activity. Indeed, we think that the dynamics of the entropy is correlated with information exchange and more generally with the building of the common ground. It becomes therefore possible to start studying the brain basis of mutual understanding by looking specifically at the brain signal associated with entropy peaks. Our hypothesis is that this entropy-based indicator could offer the possibility to analyze the brain signal in a time-locked event paradigm (evoked-related potentials) as well as the

time-frequency level (frequency bands).

Acknowledgements

This work, carried out within the Institut Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government, managed by the French National Agency for Research (ANR).

References

Mary Amoyal. 2018. Analyse du sourire lors des transitions thématiques dans la conversation.

Mary Amoyal and Béatrice Priego-Valverde. 2019. Smiling for negotiating topic transitions in french conversation. In *GESPIN-Gesture and Speech in Interaction*.

Mary Amoyal, Béatrice Priego-Valverde, and Stéphane Rauzy. 2020. PACO : A corpus to analyze the impact of common ground in spontaneous face-to-face interaction. In *LREC procs*.

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

Frederic Bechet, Benjamin Maza, Nicolas Bigouroux, Thierry Bazillon, Marc El-Beze, Renato De Mori, and Eric Arbillot. 2012. Decoda: a call-centre human-human spoken conversation corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1343–1347.

Brigitte Bigi. 2012. Sppas: un outil« user-friendly» pour l'alignement texte/son (sppas: a tool to perform text/speech alignement)[in french]. In *JEP-TALN-RECITAL 2012, Workshop DEGELS 2012: Défi GEste Langue des Signes (DEGELS 2012: Gestures and Sign Language Challenge)*, pages 85–92.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The erp response to the amount of information conveyed by words in sentences. *Brain and language*, 140:1–11.

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*, pages 65–72.

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*, pages 647–660, Online. Association for Computational Linguistics.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2021. Is Information Density Uniform in Task-Oriented Dialogues? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8271–8283, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceeding of 2nd Conference of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

Marti A Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

T Jaeger and Roger Levy. 2006. Speakers optimize information density through syntactic reduction. volume 19.

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*, pages 317–324.

James A Michaelov, Megan D Bardolph, Seana Coulson, and Benjamin K Bergen. 2021. Different kinds of cognitive plausibility: why are transformers better than rnns at predicting n400 amplitude? *arXiv preprint arXiv:2107.09648*.

Irene Fernandez Monsalve, Stefan L. Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 398–408. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Martin Pickering and Simon Garrod. 2021. *Understanding Dialogue*. Cambridge University Press.

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

Béatrice Priego-Valverde, Brigitte Bigi, and Mary Amoyal. 2020. "cheese!": a corpus of face-to-face french interactions. a case study for analyzing smiling and conversational humor. In *LREC*, pages 467–475.

Ting Qian and T Florian Jaeger. 2011. Topic shift in efficient discourse production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Tim Cistac, Pierand Rault, Rémi Louf, Joe towicz, Morgan Funand Davison, Sam Shleifer, Patrick Platen, Clara Ma, Yacine Jernite, Julien Plu, Teven Le Xu, Canand 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. 2016. Entropy Converges Between Dialogue Participants: Explanations from an Information-Theoretic Perspective. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 537–546, Berlin, Germany. Association for Computational Linguistics.

Yang Xu and David Reitter. 2017. Spectral Analysis of Information Density in Dialogue Predicts Collaborative Task Performance. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 623–633, Vancouver, Canada. Association for Computational Linguistics.

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

A Corpus Excerpt

Table 6 shows an excerpt of a Paco-Cheese dialogue, annotated with utterance position in the dialogue, current discussion theme, speaker identifiers and information content estimates (contextualised, decontextualised, and difference between the two).

B Language Models Training and comparison

B.1 Transformers

We experiment with GPT-2 (Radford et al., 2019), an autoregressive Transformer-based (Vaswani et al., 2017) language model, relying on HuggingFace's implementation, pretrained models[4] and default tokenizers.

Considering the corpora peculiarity (dialogue) which differs from most of the training data, we finetune the models on 70% of the target corpora. The finetuned models yield significantly lower perplexity on the portion of the dataset reserved for testing. One epoch with a training loss of 5e-05 (default) and batches of size 8 leads to significant improvement on the English corpus. The French model is finetuned for 20 epochs with a learning rate of 1e-05 and batches of size 16.

For inference, the model's maximum sequence length is used (1024) so as to maximize the model's ability to extract context from the discourse.

To match the SRILM execution output as well as to give context to the prediction of the first sentence token, we include a sentence beginning token at the start of the sentence for the prediction, but this token's information content is not computed.

B.2 Other language models

RNN models Data in input of the RNN models is parsed using the same Tokenizers as GPT in order to facilitate comparison between models; the models are trained on the same fraction of the corpus. After a first pass on a set of wikipedia data, the model are finetuned for 2 epochs on the target dataset. The model's architecture is as follows: one embeddings layer, one GRU layer (`hidden_size`=128). The RNN cell output is then fed to a Linear layer through a Dropout layer.

SRILM Language Models Unlike neural network models which training relied on tokenizers which virtually removed the problem of out of vocabulary (OOV) tokens, SRILM Language Models can only rely on the vocabulary encountered during training for inference of probabilities. Choosing the model therefore involves balancing perplexity and number of OOV tokens matched during inference. The fraction of OOV in the held-out data

[4]Pretrained model used for English corpora was the default `gpt2` weights; for French corpora, weights from `dbddv01/gpt2-french-small` were used.

Table 6: 20 lines from the Paco-Cheese corpora, excerpt of the conversation between AA and OR.

index	theme	speaker	text	H(S|C)	H(S)	MI
120	exams	OR	ça venait de la psycho de l'anthropo enfin de plein de euh domaines	1.18	1.30	0.13
121	exams	AA	ouais ouais	0.31	0.43	0.12
122	exams	OR	je pense c'était juste simplement euh ça	1.04	1.01	-0.04
123	exams	AA	ben ouais mais moi le truc c'est que genre la veille ben du coup je l'avais revu et tout	0.76	0.77	0.01
124	exams	OR	et	0.40	0.62	0.21
125	exams	AA	genre les j'ai vu mes deux résumés je les ai regardés j'ai fait euh	0.95	0.91	-0.03
126	exams	AA	pouah c'est bon ça tombera non non j'ai fait non c'est bon ça tombera pas sur ça	0.76	0.74	-0.01
127	exams	OR	la flemme	1.38	1.03	-0.35
128	exams	OR	ouais	0.37	0.65	0.28
129	exams	AA	genre du coup je les ai lu vite fait en diagonale	1.51	1.55	0.04
130	exams	AA	et	0.51	0.62	0.11
131	exams	AA	après j'ai eu la première question du partiel j'ai fait	0.80	0.85	0.06
132	exams	AA	ah	0.74	0.79	0.05
133	exams	OR	ouais voilà	0.65	0.65	-0.00
134	exams	AA	ah bon	1.41	1.12	-0.29
135	exams	OR	et moi je m'étais même pas rendue compte que c'était là-dedans c'est après cristèle elle m'a dit mais tu vois que c'est tu as exactement ton résumé genre	0.80	0.88	0.08
136	exams	AA	j'aurais dû	0.78	0.75	-0.03
137	exams	OR	alors qu'en plus le résumé quand elle a corrigé il m'a dit très bon résumé	1.11	1.16	0.06
138	exams	AA	ouais moi aussi	0.72	0.76	0.04
139	exams	AA	du coup j'étais un petit peu deg quoi	1.77	1.57	-0.20

was between 1 and 5% with non-finetuned models, lower with finetuned models. Following Xu and Reitter (2017) who train their language model on a different corpus, we compare different data sources for the language model. We find that pretraining the model on a larger dialogue corpus (we use DECODA, (Bechet et al., 2012)) then finetuning it on a fraction of the target corpus yields the best balance in terms of perplexity and number of OOV tokens. Indeed perplexity will be lower with large corpus that are closer in structure to the target data; thus training on dialogue data will be better than training on written corpus such as wikipedia, especially considering that the larger the original corpus, the smaller the effect of finetuning.

B.3 Building up contextual information

One interrogation that came with using models with context was how context buildup allowed for better expectations of the upcoming words. The mind is capable of selecting relevant information from an utterance and reusing it long distance, albeit with limits, as the memory span is not infinite. How much pull would long distance information have in the predictions? The biggest information input happens with the addition of the previous sentence to the context (see Figure 5); further additions to the context have a more limited impact. Thus computed values of entropy for each sentence can

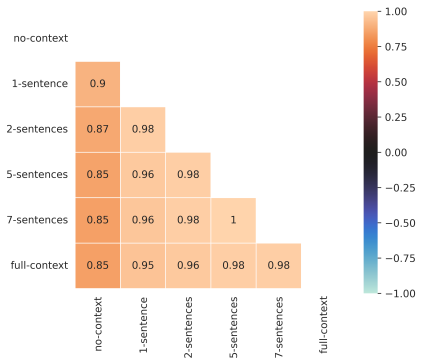

Figure 5: Correlation between entropy values given by the model, depending on the length of the contextual information, in IPUs

mostly be explained by language understanding and local structure of the sentence, with previous utterances and long distance information selection refining the predictions.

C Experimental Results

C.1 Linear Models results

Table 7 summarise the results of our statistical analysis. The same four linear models are fitted on information content estimated on different sets of the data: the utterance column refers to the length of the context, the IPU being the main condition, and `concat 1s` referring to paradigms where IPUs from one speaker are aggregated as long as they

Table 7: Results of linear mixed-effect models on the Paco-Cheese dataset

				whole dialogues			free conversation only		
				estimate	p	sig	estimate	p	sig
model	utterance	test_label	var						
H(S|C)	IPU	Position in INTERACTION	(Intercept)	-0.269	0.000	***	-0.009	0.863	
			logp	0.027	0.000	***	-0.015	0.111	
		Position in THEME	(Intercept)	-0.023	0.174		-0.010	0.577	
			logt	-0.032	0.000	***	-0.030	0.000	***
		Position in THEME - initiator	(Intercept)	-0.056	0.017	*	-0.010	0.668	
			logt	-0.021	0.002	**	-0.024	0.001	***
		Position in THEME - responder	(Intercept)	-0.082	0.032	*	-0.087	0.023	*
			logt	-0.015	0.114		-0.014	0.153	
	concat 1s	Position in INTERACTION	(Intercept)	-0.004	0.924				
			logp	-0.022	0.005	**			
		Position in THEME	(Intercept)	-0.016	0.570				
			logt	-0.029	0.000	***			
		Position in THEME - initiator	(Intercept)	0.003	0.891				
			logt	-0.033	0.002	**			
		Position in THEME - responder	(Intercept)	-0.082	0.059	.			
			logt	-0.017	0.123				
H(S)	IPU	Position in INTERACTION	(Intercept)	-0.112	0.000	***			
			logp	0.010	0.004	**			
		Position in THEME	(Intercept)	-0.011	0.344				
			logt	-0.015	0.000	***			
		Position in THEME - initiator	(Intercept)	-0.018	0.213				
			logt	-0.012	0.015	*			
		Position in THEME - responder	(Intercept)	-0.053	0.028	*			
			logt	-0.004	0.534				

are not interrupted by pauses longer than 1 second and are not interrupted by the other speaker. *Whole dialogue* and *free conversation only* refer to whether the dialogue data is considered as a whole or whether the start of the dialogue (introductions, reading of jokes to kickstart conversation) is removed only to keep the free flowing conversation. In those models, the logarithm of the information content is the response variable (H(S|C) or H(S)) and the logarithm of turn position (whether global, `logp`, or relative to the local theme, `logt`) is the fixed effect. Dialogues are considered a random effect in this analysis.

All models yield similar results in terms of estimates and p-value for the 4 conditions, with the exception of the effect of position in interaction that disappears in the free conversations only condition.

C.2 Peaks and Theme Change Locations

Manual annotation is compared to automated annotation based on lexical similarity using the TextTiling algorithm (Hearst, 1997). Figure 6 shows the distribution of annotated themes throughout two example conversations, with dashed lines indicated the start of new themes as annotated manually and automatically. TextTiling shows a higher sensitivity than human annotation to lexical changes in the conversation, resulting in a number of annotated themes twice as large on average.

Peak detection is run using two methods. The first method (LocalOutlier in the table) relies on the implementation of Local Outlier Factor by `scikit-learn` (Pedregosa et al., 2011), which allows for comparison of a value to its neighbors (n=5) to detect locally unusual values. The second method (NormOutlier) relies on a global, where only the top 2% values are considered outliers (see Figure 7). Both methods are applied to series of contextualised entropy H(S|C) as well as mutual information (MI) as both would be expected to be sensitive to the introduction of new information to the conversation. Neighbors number and percentage threshold value were chosen as optimal values based on accuracy, precision and recall, on a subset of the data.

Table 8 summarises how peak location and TextTiling theme break prediction fare in predicting the location of manually annotated theme changes. There is a total of 268 of theme changes in the dataset (excluding moments annotated as transitions between two themes). We consider that the location of a theme change might not be an accurate consideration since it depends on the annotator sensitivity and consider the prediction might match a location within a small window of IPUs centered around it. Windows of size 2 and 5 were consid-

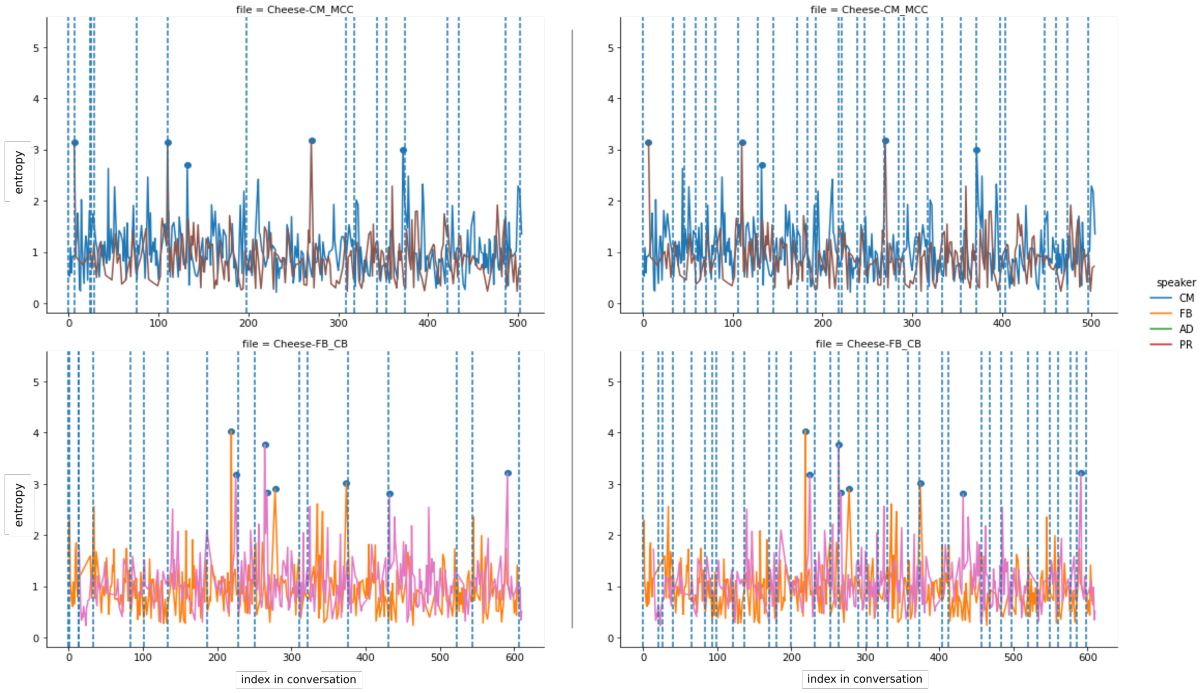

Figure 6: Evolution of normalised contextualised entropy on two dialogues. The two speakers are plotted in different colors. Bold points indicate outliers detected by the NormOutlier method. Dashed lines indicate the start of new themes - left: manual annotation; right: predicted by TextTiling

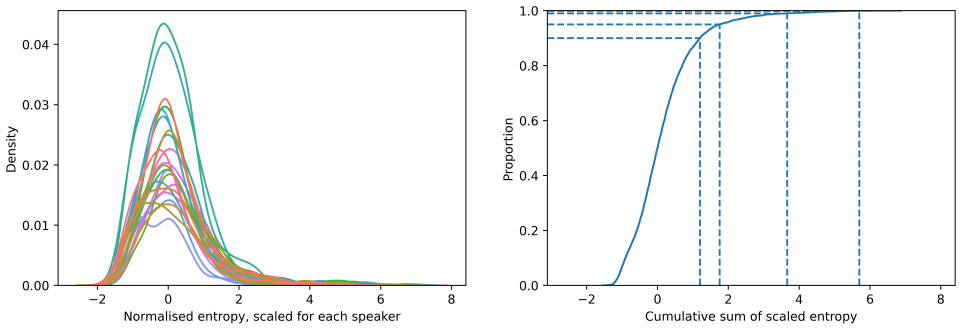

Figure 7: Probability distribution function of entropy (normalised and scaled) observed individually for the various speakers in the corpora (left) and cumulative (right), with ticks at 90, 95 and 99% density.

ered (larger windows were discarded as themes can change frequently). Table 9 compares the location of peaks and that of lexical changes as annotated by the TextTiling algorithm. In both tables, precision and recall refer to commonly used metrics counting the number of exact prediction (True Positives) compared to the number of peaks that weren't located at theme breaks (False Positives) and theme breaks which did not result in entropy peaks (False Negatives).

LocalOutlier systematically yields a larger number of locations identified as peaks whereas NormOutlier is more sparse - which was expected by design of NormOutlier. Focusing LocalOutlier on outliers that increase entropy does not improve prediction. Both algorithms detect a smaller number of locations of interest when only taking into account evolution of the contextualised entropy ($H(S|C)$) over the dialogue rather than mutual information. TextTiling is not more accurate than peak detection to detect manually annotated theme change locations, but a larger number of peaks matched theme change predicted by TextTiling. However none of those methods perform better than a baseline classifier (stratified Dummy Classifier from the `scikit-learn` library) trained to predict boundaries of thematic episodes based on values for $H(S|C)$, MI and utterance length.

Increasing the window size reliably increases the number of theme breaks matched, substanti-

ating the hypothesis that theme changes involves adding new information to the conversation, which is detectable using entropy metrics.

D In-depth per-word entropy analysis

D.1 Choice of words and peaks of entropy in a discussion

We select the sentences which have been labelled as peaks of entropy and analyse entropy word by word, how each word contributes to the sentence, and what category those words fall into. Examples of these words are given in Table 10. What we find is that most words with high entropy are simply rare enough that they are deemed improbable; most of them are either nouns (48%), adjectives (20%) or verbs. Some peaks however are caused by words falling into one of the following categories (disregarding words from the transcription that still contain typos): Proper Nouns (*Arthur, Jas, Danemark, Luminy...*), contractions or abreviations (*FLE, QCM*), technical words that might be taken from other languages (*slides, rift...*) which would be highly unusual for the model, but part of the shared knowledge for the two interlocutors.

The more an unexpected word is linked to a theme, the more we would expect it to reappear, and if it had caused a peak of entropy at first, we would expect that surprise being smoothed over time. Indeed, on a conversational level, the more a word occurs in conversation, it becomes part of the shared knowledge and is expected to be reused by any locutor. As a consequence, reused references are subject to compression throughout a dialogue (Giulianelli and Fernández, 2021) as they are expected to be understood without much cognitive load the more they appear. Context (previous words mentioned in the conversation) being available to our models they should equally be able to not be surprised by the reappearance. In our case, most words causing peaks are not reoccurring (68%), but those that do indeed become slightly more predictable (generating slightly less entropy, $p < 0.1$)

D.2 The role of backchannels

Backchannels are words or movements (nods, smiles) that a listener will spontaneously produce to signal the speaker of their attention, encourage them to continue with their story or on the contrary signal their lack of understanding or disagreement. Several kinds of feedbacks are annotated in Paco-Cheese, based on speech production, nods, smiles and context: generic (*hm, yes, ok, sure...*) and specific (context-dependant productions, whether positive or negative).

Considering that some feedback productions seemed to appear in the list of peaks, we analyzed in more details how well the models - which were initially designed for written language, devoid of backchannels - adjust to such phenomena after fine-tuning. A supposition was that feedbacks might appear as "disruptive" in the written flow of conversation, since productions are often partial or context-dependent.

We expected generic feedbacks to be well adapted to; specific feedbacks however would be contextual and generate slightly more entropy. Indeed, productions labelled as generic feedback are associated with per-sentence entropy values that are lower ($p < 0.01$) that those of productions that do not contain feedbacks. Specific feedbacks are associated with higher entropy values than generic feedbacks, but in the majority of cases (negative-unexpected feedbacks excepted) associated with lower entropy values than the productions not containing any feedback ($p < 0.05$) (see Table 11).

Table 8: Comparing TextTiling theme change locations and information content peaks to manually annotated theme changes. Baseline classifier (DumStrat) is added for consideration. TP indicates the number of elements, that either directly match a manual annotation or fall within a small window of that point.

			exact prediction			window=2			window=5		
input data	algorithm	nb elements	TP	precision	recall	TP	precision	recall	TP	precision	recall
text	TextTiling	565	28	0.050	0.104	28	0.060	0.108	80	0.136	0.299
MI	LocalOutlier	2137	71	0.033	0.265	118	0.066	0.440	193	0.154	0.720
	NormOutlier	70	3	0.043	0.011	3	0.086	0.011	8	0.129	0.030
H(S|C)	LocalOutlier	1602	59	0.037	0.220	101	0.072	0.377	173	0.172	0.646
	NormOutlier	138	6	0.043	0.022	9	0.101	0.034	23	0.174	0.086
	DumStrat	261	15	0.057	0.027	28	0.115	0.050	64	0.268	0.113

Table 9: Comparing information content peaks to the locations of TextTiling theme changes. TP indicates the number of elements, that either directly match a manual annotation or fall within a small window of that point.

			exact prediction			window=2			window=5		
input data	algorithm	number of elements	TP	precision	recall	TP	precision	recall	TP	precision	recall
H(S|C)	LocalOutlier	1602	112	0.070	0.198	162	0.125	0.287	318	0.280	0.563
	NormOutlier	138	15	0.109	0.027	14	0.159	0.025	31	0.225	0.055
MI	LocalOutlier	2137	122	0.057	0.216	203	0.117	0.359	381	0.278	0.674
	NormOutlier	70	1	0.014	0.002	2	0.071	0.004	14	0.200	0.025

Table 10: Words with the highest entropy that appear in utterances labelled as peaks

'arthur', 'improbable', 'rift', 'interagir', 'mesuré', 'aram', 'anthropologie', 'jugé', 'jas', 'autes', 'deg', 'opposés', 'ent', 'moinl', 'laide', 'pas', 'identifie', 'quarantaine', 'danemark', 'audience', 'ets', 'saint', 'conte', 'sû', 'comparent', 'qcm', 'coup', 'implicite', 'anonyme', 'explicite', 'dis', 'calédonie', 'didons', 'tain', 'maléfique', 'géologie', 'dirigés', 'exemp', 'londres', 'craintes', 'médhia', 'incompréhension', 'montrer', 'décennie', 'ydis', 'dit', 'tien', 'règles', 'temps', 'cont', 'pt', 'dénonce', 'allée', 'devoirs', 'discours', 'là', 'fle', 'vêtement', 'cing', 'lie', 'occupé', 'anova', 'emmener', 'énorme', 'suppose', 'bianca', 'trois', 'humoristique', 'obliger', 'professeur', 'particuliers', 'sociale', 'oculus', 'totallement', 'alcooliques', 'la', 'bas', 'intro', 'teint', 'techniquement', 'régression', 'suisse', 'interêt', 'luminy', 'clés', 'quantité', 'perspective', 'morphologie', 'vive', 'istres', 'smaines', 'cognitive', 'contraignant', 'stricto sensu', 'afrique', 'occupe', 'pénal', 'voyage', 'apprécies', 'psychologue'

Table 11: Number of feedbacks of each category in the corpus and length compared to that of productions that don't contain feedbacks

Production type	# occurrences	average length	average entropy	comparison (t.test) of entropy: pvalue	
				less than 'no-feedback'	more than 'generic'
no-feedback		8.2	1.084 ± 0.51		<0.001
generic	799	2.0	0.651 ± 0.41	<0.001	
négative-expected	339	4.4	0.920 ± 0.54	<0.001	<0.001
négative-unexpected	303	4.2	1.055 ± 0.55	0.32	<0.001
positive-expected	110	4.7	1.010 ± 0.60	0.01	<0.001
positive-unexpected	75	3.7	0.983 ± 0.55	<0.001	<0.001