Locating strongly informative utterances in conversation using multimodal cues

Eliot Maës, Philippe Blache, Leonor Becerra-Bonache

Locating strongly informative utterances in conversation using multimodal cues

Eliot Maës (eliot.maes@lis-lab.fr)
Aix Marseille Univ, CNRS, LIS,
Marseille, France

Philippe Blache (philippe.blache@univ-amu.fr)
Aix Marseille Univ, CNRS, LPL,
Marseille, France

Leonor Becerra-Bonache (leonor.becerra@lis-lab.fr)
Aix Marseille Univ, CNRS, LIS,
Marseille, France

Abstract

Interaction theories argue that mutual understanding between speakers in natural conversations arises from building shared knowledge (*common ground*), but no model specifies what information is retained or under what conditions. Previous studies have used Information Theory metrics to quantify the dynamics of information exchanged between participants but lack an efficient way to identify which information becomes common ground. These attempts furthermore limited themselves to the study of conversation transcripts, overlooking nonverbal cues like visuals and intonation. To address this, we propose a method for annotating new corpora using models trained on a subset of annotated utterances. Results show a fair applicability ($\kappa \simeq 0.3$) across corpora, though this is strongly modulated by the conversational task being investigated.

Keywords: informativity, multimodal learning, dialog act, prediction

Introduction

Humans' ability to understand each other despite differences in experience and personality remains a challenge to model in detail. Grounding theory holds that a contribution to dialogue is only successful once the interlocutors demonstrate mutual evidence of understanding (Clark, 1996). Therefore, successful communication depends on efficiently sharing information with others in ways that ensure being understood. Interaction theories link this ability to the gradual alignment of linguistic representations between speakers (M. Pickering & Garrod, 2021; M. J. Pickering & Garrod, 2004). In these frameworks, all conversational tasks, whether they involve collaboration or competition, rely on the speakers' ability to build a set of shared knowledge (also known as *common ground*); the higher their ability to align on such knowledge and develop their understanding, the deeper and higher quality the interaction, and the stronger the convergence at lexical, syntactic, prosodic, gestural, and behavioral levels.

However, no comprehensive model explains how and under what conditions information enters common ground. One hypothesis is that the convergence phenomenon - and especially its semantic component - indicates a specific coordination between participants in terms of information exchange, which can be further analyzed by studying the amount of exchanged information and its dynamics during a conversation. Previous studies have looked at information transfer through Dialog Act annotation or Information Theory metrics (Shannon, 1948). However, these do not provide with an efficient way to *study individual exchanges or quantify the amount of information transferred by an utterance*. Building on Clark (1996) theory of grounding, we take the view that the novelty of an information is not the only thing that matters for inclusion in the common ground; other factors, such as information understanding and speakers' relative interest for the topic at hand, will also affect this process.

The present study contributes five elements toward that goal. (1) We formalise strongly informative utterances as a four-level scale of Information Content (LEV) that unifies insights from information-status tagging and dialog-act research. (2) We annotate a subset of two multimodal French corpora with LEV and dialog-act labels, including both free conversation tasks and a conversational task. (3) We investigate the impact of each modality (audio, video, text) in the modeling of this content using transformer encoders and recurrent networks. (4) We benchmark a model fusing lexical, prosodic and visual cues, showing that multimodality yields consistent, if modest, gains. (5) We open a path for cognitive scientists to correlate high-LEV windows with downstream behavioural or neural measures, thereby investigating theories of grounding.

The paper is organized as follows. We first situate our research in its context, with our reasons for undertaking these experiments. The next section is dedicated to the developed annotation scheme and the corpora it was applied to. An account of the models used for the prediction is then given, before detailing our results and the conclusions drawn from them.

Previous Works

Studying Common Ground in Conversation

The complexity in the study of common ground development arises from the inherently dynamic nature of dialogs. Convention and assumptions are rarely imposed by just one speaker, most pieces of information being added upon by all interlocutors, co-constructing utterances. Co-occurring utterances may also refer to different pieces of information, with their grounding therefore occurring out of order. In addition, interlocutors do not always manifest overt signs of having grounded.

For this reason, most studies on common ground focus on conversational tasks - rather than free conversation - that allow for simpler dialogue with more limited and controlled

vocabulary, yet higher engagement making participants strive for mutual understanding. Investigations also usually focus on one specific phenomena - conversational feedbacks (Clark and Brennan (1991) or Roque and Traum (2008), operationalising common ground through evidence of acceptance), repetitions, speech dynamics (Bertrand & Espesser, 2017), speaker beliefs (Markowska et al., 2023)... - with limited investigation of how these different features interact together.

Informativity in Conversation

A direct study of information transfers in conversation could give another angle on the development of the common ground. Indeed, despite most utterances serving some informative role, not everything shared has a direct impact on the shared knowledge between speakers: some information is not as interesting to a speaker as it is to the other, and will be forgotten upon hearing; in other cases, a larger buildup of information is required in order for a point to be understood, leading to a "eureka" moment.

Various approaches have been investigated to quantify and analyze how speakers share information. Purely based off linguistics features, feedback mechanisms provide insight into information uptake, as listener reactions signal whether an utterance meaningfully updates the common ground. Dialogue act classification, while useful in identifying informative utterances ("inform" or "statement" label), remains broad and often task-specific, making it less suited for open-ended conversations (Klein et al., 1998). Kravtchenko and Demberg (2022) noticed that utterances that are informationally redundant can also trigger pragmatic inferences. Finally, nonverbal cues can modulate perceived importance(Roettger et al., 2019) or help locate "hot-spots" of informative content in dialog (Wrede & Shriberg, 2003).

Another angle into this topic is uses Information Theory, relying on the correlation between the surprisal of a word in a sentence and the cognitive load associated with its parsing to introduce an utterance-based entropy metric (Giulianelli et al., 2021; Xu & Reitter, 2018) to explore information patterns at the scale of a conversation. The finer results of this metric are however highly dependent on the training of a given model, which makes their use adequate for exploring general patterns, but less so for the exploration of unique or more localised phenomena. Conversation summarization (Maës et al., 2024) offers another perspective, as the process of summarization implicitly selects the most informative content, mirroring speakers' real-time evaluation of relevance.

Khebour et al. (2024) also reports on training a model for Common Ground Tracking model to identify both the current set of beliefs and the associated level of evidence, using a data structure to differentiate between facts and evidence known to the group, and topics that are being discussed.

Multimodal Learning

Until recently, multimodal exploration was limited to easy to model phenomena with large corpora available for training models, such as emotion prediction (Kalateh et al., 2024; Ramaswamy & Palaniswamy, 2024) in conversation. The development of more foundation models and widens the scope of problems which can be investigated, though the question of the adequate handling of these features remains a question: how can we best highlight the co-occurrences of events across the channels.

Corpus

We focus on two corpora for the development of our method: the BrainKT corpus (Maës et al., 2023) which contains multimodal data including brain activity - which will enable the future investigation of cognitive processes behind production and reception of information - and the PACO-CHEESE dataset, which was partially annotated for informativity content in previous experiments (Maës et al., 2024).

BrainKT is a multimodal corpus containing the audio, video and neurophysiological (EEG, heart rate, skin conductance) recordings of 28 interactions between dyads of participants talking through two conversational tasks. Conversations are in French and lasting about 30 minutes (15 minutes for each task). Through the tasks, the participants would familiarize themselves with one another, build common ground and use increasingly unconstrained vocabulary and references. The first task is a collaborative game; after a short dilemma serving as the discussion prompt, they were then free to discuss the topics of their choice.

PACO-CHEESE is a multimodal corpus containing audio and video recordings of 26 interactions between dyads of participants. Conversations are in French and last between 15 and 20 minutes. Participants were given a short prompt to read (*donkey joke, frog joke*) to elicit conversation before continuing the talk on the topics of their choice. For 16 out of the 26 recordings, participants were not acquainted. The corpus is furthermore enriched with annotations for noise, laugh, pauses, feedbacks, head nods and smiles (Amoyal, 2018; Amoyal & Priego-Valverde, 2019). Expert thematic annotation has been added to 16 of the dialogues.

For both corpora, the manual transcription was aligned to the audio signal. Consequently, the speech segments we consider here are utterances or inter-pausal units (IPUs), segments of speech of which boundaries are defined by pauses longer than 200ms of silence.

Informativity Annotation

The working assumption for the current study is that information transfers may occur continuously throughout conversations but the quantity of information being conveyed and the impact on the conversation varies, with for instance positive feedback amounting to a negligeable amount of information compared to the answer to an open-ended question. Those utterances are more likely to have an impact on the

cognitive processes involved in common ground building. We take inspiration from existing studies looking into common ground and information annotation. The emergence of the common ground can be described at the scale of the conversation through the progressive alignment between speakers. It is more however most often described at smaller scales, such as one or several neighboring utterances, or even word-level. For instance, testing language models' ability to locate events or information uses the frame annotation. Turn-Taking and language prediction investigation are two other cases using word-level annotation, with Bögels et al. (2015) warranting this requirement by noting that participants answering quiz questions for which response planning could start early would do so. At the utterance level, Dialog Acts are a common occurence in dialogue analysis; they can be used (for instance) to annotate and investigate dialogue structure and speaker intent. The belief-based annotation used by Markowska et al. (2023) details events mentioned (uttered or inferred) throughout dialogue utterances and whether and how strongly each speaker believes them; this is a rare instance of a scheme keeping track both speakers' knowledge in parallel, though this requires a higher level of implication from annotators who are required to postulate the internal representations of the speakers. Finally, Common Ground Units described by Nakatani and Traum (1999) are groups composed of a few utterances that may or may not be adjacent and are relative to the exchange of one information.

The information transfers we are targeting are thus those which are going to steer and organise the conversation. They have a mostly local scope - they may be recalled or have an impact later in the conversation, but they are agreed on right after being mentioned. Previous experiments Maës et al. (2024) however showed the difficulty in describing the impact, in terms of informativity, of utterances. Supplementary labeling was thus added to better describe the conversation, with a small set of Dialog Acts (see Table 2) and a word-scale annotation for words being empathized by a speaker or carrying most of the expected information. Considering the schemes have to fit both free conversation and game tasks, we added a label to separate utterances where the speaker is reading instruction out aloud rather than mindfully sharing information. Table 1 lists the labels used for this Informativity Content (LEV) annotation.

Inter-annotator agreement and annotation validity A first round of annotation led to the simplification of the original set of labels chosen for the Dialog Act annotation, as the inter-annotator agreement (computed on one file) was quite low, with cohen's kappa score: $\kappa_{DA} = 0.45$. The agreement for informativity content was lower ($\kappa_{LEV} = 0.27$) but not unexpected. As a reference, κ_{LEV} on the PACO-CHEESE annotation was also quite low on average (0.326 ± 0.101, reaching 0.493 for only one pair of annotators), despite focusing on selected themes of natural conversations rather than complete dialogues. The agreement was higher on DA labels on the game subset of the BrainKT corpus, but higher on LEV labels on the free conversation part - as the more interesting utterances are more obvious in free conversation.

Level	Description	Percent.
1	The IPU conveys *major information*, strongly correlated to the currently discussed theme, which has *never been mentioned* before	7.61%
2	The IPU conveys information, but of *secondary importance* to the conversation; it can include utterances that detail previously mentioned information	18.96%
3	The IPU conveys little to *no information*, or a repetition of aforementioned information	73.43%
4	The speaker is reading experiment guidelines out loud	

Table 1: Description of labels used in the Information Content (LEV) annotation

General Label	Sub categories	Description	Percent.
Statement	ST-New	Facts useful to the speaker (new)	40.0%
	ST-Expl	Facts useful to the speaker (detailed explanations, repetitions)	
Feedback	FDBK	Agree / Reject / Repetitons / sentence completion / non understanding	20.9%
	FDBK-Spe	Feedback specific to the previous utterance	
Instruct	Instr	Directing or guiding the listener to perform an action	5.5%
Others	Oth	Rhetorical, self-talk, uninterpretable, interrupted utterances	19.6%
	Oth-Opin	Opinions, emotions	
	Oth-Q	(Y/N, Wh-) Questions	9.5%
	Oth-Ans	Yes/No answers	4.3%

Table 2: Description of labels used in the Dialog Act (DA) annotation

Considering the cost of human annotation, we explored the reason for this lower agreement. Recent work into annotation aggregation has highlighted that annotator disagreement may not always be due to errors, but rather to differences in interpretation (which is however not logged during the process), and that models would benefit from learning from these various sources rather than from an aggregated view (Mokhberian et al., 2023; Plank, 2022; Weber-Genzel et al., 2024).

Upon review, it was however noted that one annotator misunderstanding of some of the annotation guidelines was leading many of the disagreements (for instance what constituted *answers* to questions *vs.* statements or feedbacks), which led us to only keep one set of annotations for the experiments.

Models and Experiments

We train models to predict both the annotated Dialogue Acts and Informativity levels in parallel. We hypothesize that this might help the model determine whether an utterance is important for the conversation or not. Models are either unimodal (text or audio or video) or multimodal (text-audio or all 3). The prediction is done at the end of each utterance, using utterance content (tokenized) and / or audio and video signal, cut and 0-padded to a 2s-window. (adding more signal didn't not significantly improve the prediction, neither did adding signal after the utterance end; on the other end, cutting the window to 1 or 1.5s significantly decreased performances).

Text and audio models were adapted from HuggingFace's `transformer` (Wolf et al., 2019) SequenceClassifier models with separate Linear layers for each target. Models were trained using the `transformer` Trainer with a custom loss function, computing the sum of the `CrossEntropyLoss` with adequate class weights for each target. Batches contained a balanced number of samples for each target. 4-cross validation was used to test the stability of the predictions.

Text Embeddings For unimodal models, we compared the performances of two encoder models (FlauBERT (Le et al., 2020), DeBERTa (Antoun et al., 2023)) and one decoder model (DialoGPT[1] (Zhang, 2019)) for the classification both of LEV and DA values. For multimodal models, only the DeBERTa embeddings were kept.

Audio Embeddings We chose to use `wav2vec2` (Baevski et al., 2020) embeddings for the audio signal. Pretrained weights were taken from either Parcollet et al. (2023) (base model, 12 hidden layers) and Grosman (2021) (large model, 24 hidden layers). Rather than using the last layer's hidden state for the classification, we used learnable parameters to get the optimal weight distribution for the weighted average of the information from the hidden layers.

Video Representation We trained RNN-based models on OpenFace data (features are described in Table 3) which are commonly used in this kind of task. We especially focused on eye and action unit features.

Modality Fusion For models using 2 or 3 modalities, we separately obtain embedding representations for each modality then concatenate them before the classification step. Weights were initialized using pretrained models for the text and audio modalities, and randomly initialized for the video (see Figure 1).

Label distribution Due to the lack of representation and difficulty to identify some categories, only 6 (out of 9) DA labels and the first 3 LEV labels were kept. (Groups are indicated in the percentage column in Tables 2 and 1).

[1] models finetuned from `emil2000/dialogpt-for-french-language`

Features	Number of features	Description
gaze_	8	
eye_lmk_	280	eye landmarks (gaze details)
pose_	6	
x_, y_,	124x2	landmarks in 2D
X_,Y_, Z_,	124x3	landmarks in 3D
p_	40	Rigid and non-rigid shape parameters - identity
AU_r	17	Facial Action Units (intensity)
AU_c	18	Facial Action Units (exists)

Table 3: Description and number of OpenFace features

Weights were applied to the loss to take this imbalance into account, with experiments showing that using inverse proportion of labels compared to the most common label ($\max(\{p\})/p(label)$, making $weights_{DA} \simeq [1,2,8,2,4,9]$ for labels in the order given the table) got close to optimal results. Increasing the weights for underrepresented samples more than that had a negative impact on the performance of the more common labels.

Results

Model results are described in Table 4. Various experiments were run to try and better understand predictions with regards to the task and the corpus subset. We use both the f_1 score and Cohen's Kappa Score κ (which takes label probabilities into account) for the evaluation.

Label Prediction

We obtain much better results for the prediction of DA labels than LEV labels, with $\kappa_{DA}simeq$ 0.52. Most DA labels are well classified, with the `Oth` and `Oth-Ans` labels being the exception to this rule - which is understandable since the model does not have access to previous utterances, making Yes/No answers and more conversational utterances difficult to distinguish from feedbacks. Adding context however did not improve the prediction.

LEV labels, on the other hand, are less easily classified ($\kappa_{LEV} \simeq 0.3$ on BrainKT, 0.4 on PACO-CHEESE), as weights included in the loss do not seem to be enough to completely offset the imbalanced learning. This is however not too far from the agreement between human annotators on the corpus, which is a good sign considering the difficulty of the task.

Models trained on the different data splits also show variations in their prediction of the labels, stressing the need for a larger dataset to train on. Model agreement (computed with Fleiss' Kappa) is fair, with an average of $\kappa_{DA} \simeq 0.69 \pm 0.7$ for the DA prediction task and $\kappa_{LEV} \simeq 0.48 \pm 0.1$ for the LEV task for the models, with similar agreement on each corpus subset. Most utterances receive the same predicted label across splits, indicating that variance is concentrated in a small subset of borderline cases. This variability however seems to improve performances for models with good enough

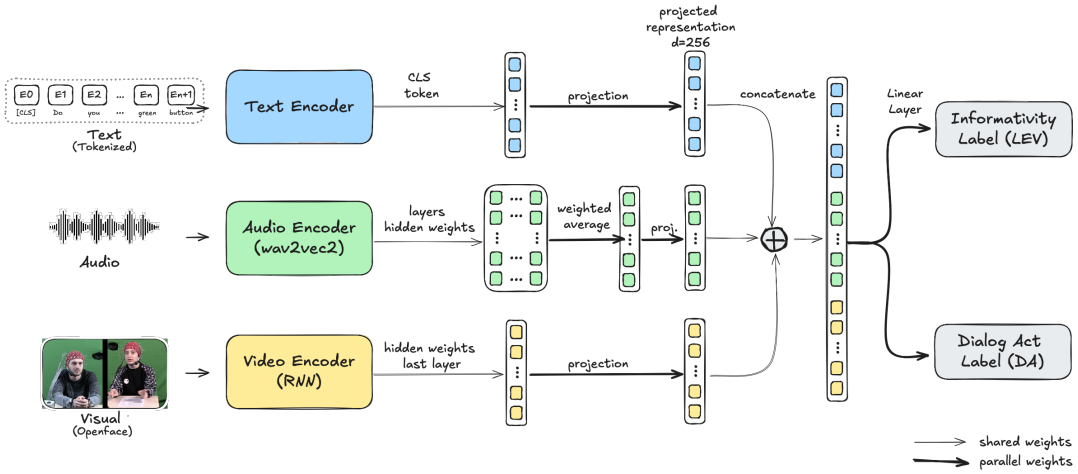

Figure 1: Description of the architecture of the multimodal model used (here with 3 modalities). A concatenation step enabled the fusion of information from the different modalities.

n	modality	pretrained	fusion type	κ_{LEV}	f^1_{LEV}	κ_{DA}
1	audio	large		0.250±0.018	0.346±0.005	0.479±0.011
1	audio	base		0.279±0.013	0.484±0.013	0.350±0.009
1	video	RNN		0.134±0.014	0.399±0.015	0.182±0.01
1	text	DeBERTa		0.276±0.028	0.476±0.015	**0.550±0.007**
2	text+audio	DeBERTa+base	concat	0.276±0.023	0.474±0.022	0.514±0.052
3	text+audio+video	DeBERTa+base+RNN	concat	**0.293±0.047**	0.477±0.022	0.525±0.08

Table 4: Summary of the main results by modality(ies) tested. In bold the model with the best results for each target. f_1 is only indicated for LEV prediction as the imbalance in labels more strongly affects the metric.

performances, as the fused prediction (most common label being selected for an utterance) leads to slightly better performances than the split models on their own ($\kappa_{LEV,agg} = 0.341$ and $\kappa_{DA,agg} = 0.607$ for the 3-modal model in Table 4).

The joint learning of LEV and DA labels does not significantly improve performances for most models - the exception being text encoder models, which benefit the most from it. With audio models, the choice of the model had a stronger impact on performances than variations in the length of the signal window chosen, with large models being better at classifying DA labels, but base models being better at LEV labels. The impact of the fusion of modalities is also limited, which can either be linked to the difficulty of the task or to the simplistic way this fusion was handled (a simple concatenation layer) which does not represent accurately enough the complex interactions between the audio and semantic contents, and the facial features.

Conversational tasks specificity

Models were trained separately on the game and free conversation corpora subsets to check on models ability to generalise to other corpora - an important question as most corpus / experiments only focus on conversational tasks. Despite both LEV and DA labels appearing in similar proportions, performances on one subset did not propagate to the other (see Figure 2). This confirms the need to study conversational tasks and free conversation in parallel to better understand how common ground building works. Game utterances appear closer to those in the free conversation than to the dilemma discussion. This might be a specificity of the BrainKT corpus as speakers were unacquainted and relied more on questions when learning about one another, whereas opinions were more frequent in the dilemma. Models performed as well on BrainKT free conversation as on PACO-CHEESE. Audio and Text models trained on PACO-CHEESE however do not perform as well on BrainKT, which could be linked to the construction of the PACO-CHEESE annotation, which only includes long themes but no theme transitions, making those utterances in other corpora difficult to predict for the model. Video models, on the other hand, perform similarly to slightly better, which would underline the benefits of learning on cherry-picked samples. One surprising result was the performances of the game-trained audio model, which performed better on free conversation corpora. This could indicate that similar prosodic patterns are used between the subtasks, but that the game subset is overall harder to analyse in terms of informativity as most utterances transfer some information.

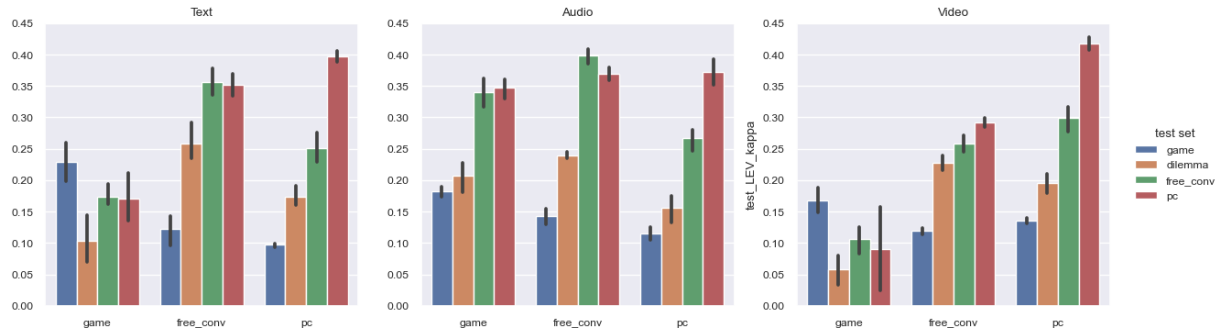

Figure 2: Performances in the prediction of the LEV labels by text models (similar trends are obtained with DA / audio and video modalities) depending on the subset of corpus on which the model was trained (game, free conversation, PACO-CHEESE).

For models trained on both subsets, DA prediction was easier on the game corpus - where utterances go more to the point - whereas LEV prediction was easier on the free conversation task - where the distinction between important information and feedbacks or anecdotes shared just to empathize a point.

Annotating the rest of the corpus

Only 7 of the BrainKT conversations were initially annotated with DA and LEV labels; in order to fully profit from the corpus, we used the best model to predict classes on the 21 remaining dialogues. A total of 1300 LEV-1 utterances are obtained throughout the corpus, out of which 66% are Statements.

LEV Label	**1**	**2**	**3**
Annot. (7)	390 (72%)	950 (75%)	14000 (27%)
Predict. (21)	910 (64%)	4750 (70%)	3700 (33%)
Whole	1300 (66%)	5700 (75%)	17700 (32%)

Table 5: Amount (approximated) of each LEV label on the whole corpus after prediction on the un-annotated dialogues. Between brackets is the fraction of utterances that are labeled as "Statement".

A manual check of the predictions however revealed a very different outlook for the two conversational tasks. In the game subset, the models relied heavily on the identification of certain keywords to label the informative content of utterances, missing the point in a large number of dialogues. The free conversation prediction is on the other hand much more reliable. Though a cherry-picking step might still be required prior to the brain activity analysis, the number of utterances available makes this prediction step a much needed gain of time.

Discussion

These experiments aimed both at exploring the developed annotation and the impact of combining modalities for the prediction of moments where information transfer is more emphasized in the conversation, with the goal of pre-annotating the BrainKT corpus for future investigation of cognitive activity during common ground building in conversations.

Though some aspects of the training (notably the limited impact of the combination of multimodal features, or the exact impact of the video features) underlines the necessity for a more in-depth analysis, some lessons can still be learned from it. The different strategies used to share information during either during free conversation *vs* conversational games have an impact on a model's (trained on any modality) ability to predict to another kind of corpus, which both validates their investigation separately and underlines the need for more parallel studies.

Overall, we have shown that grounding-oriented informativity can be detected — even if imperfectly — directly from naturalistic conversational signals. The resulting multimodal predictor already matches the lower range of human agreement and thus offers a practical filter for large corpora, opening the way for a more in-depth analysis of speaker behaviors, both with "perceivable" features (voice, facial expression, choice of words) but also of the cognitive processes in play during conversation.

References

Amoyal, M. (2018). Analyse du sourire lors des transitions thématiques dans la conversation.

Amoyal, M., & Priego-Valverde, B. (2019). Smiling for negotiating topic transitions in french conversation. *GESPIN-Gesture and Speech in Interaction*.

Antoun, W., Sagot, B., & Seddah, D. (2023). Data-efficient french language modeling with camemberta. *arXiv preprint arXiv:2306.01497*.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, *abs/2006.11477*.

Bertrand, R., & Espesser, R. (2017). Co-narration in french conversation storytelling: A quantitative insight. *Journal of Pragmatics*, *111*, 33–53.

Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific reports*, *5*(1), 12881.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association.

Giulianelli, M., Sinclair, A., & Fernández, R. (2021). Is information density uniform in task-oriented dialogues? In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 8271–8283). Association for Computational Linguistics.

Grosman, J. (2021). Fine-tuned XLSR-53 large model for speech recognition in French.

Kalateh, S., Estrada-Jimenez, L. A., Nikghadam Hojjati, S., & Barata, J. (2024). A systematic review on multimodal emotion recognition: Building blocks, current state, applications, and challenges. *IEEE Access*, *PP*, 1–1.

Khebour, I. K., Lai, K., Bradford, M., Zhu, Y., Brutti, R. A., Tam, C., Tu, J., Ibarra, B. A., Blanchard, N., Krishnaswamy, N., & Pustejovsky, J. (2024). Common ground tracking in multimodal dialogue. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 3587–3602). ELRA; ICCL.

Klein, M., Bernsen, N. O., Davies, S., Dybkjær, L., Garrido, J., Kasch, H., Mengel, A., Pirrelli, V., Poesio, M., Quazza, S., et al. (1998). Mate deliverable d1. 1: Supported coding schemes.

Kravtchenko, E., & Demberg, V. (2022). Informationally redundant utterances elicit pragmatic inferences. *Cognition*, *225*, 105159.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., & Schwab, D. (2020). Flaubert: Unsupervised language model pretraining for french. *Proceedings of The 12th Language Resources and Evaluation Conference*, 2479–2490.

Maës, E., Boudraa, H., Blache, P., & Becerra-Bonache, L. (2024). Did you get it? a zero-shot approach to locate information transfers in conversations. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 4877–4890.

Maës, E., Legou, T., Becerra, L., & Blache, P. (2023). Studying common ground instantiation using audio, video and brain behaviours: The brainkt corpus. *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, 691–702.

Markowska, M., Taghizadeh, M., Soubki, A., Mirroshandel, S., & Rambow, O. (2023). Finding common ground: Annotating and predicting common ground in spoken conversations. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: Emnlp 2023* (pp. 8221–8233). Association for Computational Linguistics.

Mokhberian, N., Marmarelis, M. G., Hopp, F. R., Basile, V., Morstatter, F., & Lerman, K. (2023). Capturing perspectives of crowdsourced annotators in subjective learning tasks. *arXiv preprint arXiv:2311.09743*.

Nakatani, C., & Traum, D. (1999). Coding discourse structure in dialogue (version 1.0).

Parcollet, T., Nguyen, H., Evain, S., Boito, M. Z., Pupier, A., Mdhaffar, S., Le, H., Alisamir, S., Tomashenko, N., Dinarelli, M., Zhang, S., Allauzen, A., Coavoux, M., Esteve, Y., Rouvier, M., Goulian, J., Lecouteux, B., Portet, F., Rossato, S., … Besacier, L. (2023). Lebenchmark 2.0: A standardized, replicable and enhanced framework for self-supervised representations of french speech.

Pickering, M., & Garrod, S. (2021). *Understanding dialogue*. Cambridge University Press.

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

Plank, B. (2022). The'problem'of human label variation: On ground truth in data, modeling and evaluation. *arXiv preprint arXiv:2211.02570*.

Ramaswamy, M. P. A., & Palaniswamy, S. (2024). Multimodal emotion recognition: A comprehensive review, trends, and challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *14*(6), e1563.

Roettger, T. B., Mahrt, T., & Cole, J. (2019). Mapping prosody onto meaning – the case of information structure in american english. *Language, Cognition and Neuroscience*, *34*(7), 841–860.

Roque, A., & Traum, D. R. (2008). Degrees of grounding based on evidence of understanding. *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*.

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

Weber-Genzel, L., Peng, S., de Marneffe, M.-C., & Plank, B. (2024). Varierr nli: Separating annotation error from human label variation. *arXiv preprint arXiv:2403.01931*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, *abs/1910.03771*.

Wrede, B., & Shriberg, E. (2003). Spotting "hot spots" in meetings: Human judgments and prosodic cues. *Proceedings of Eurospeech 2003*, 2805–2808.

Xu, Y., & Reitter, D. (2018). Information density converges in dialogue: Towards an information-theoretic model. *Cognition*, *170*, 147–163.

Zhang, Y. (2019). Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.