Evaluating Spoken Language Features in Conversational Models: The Case of Disfluencies and Feedbacks

Oussama Silem[1], **Maïwenn Fleig**[3,4,5], **Houda Oufaida**[1], **Leonor Becerra**[3,5], **Philippe Blache**[4,5]
Ecole nationale Supérieure d'Informatique (ESI), Algiers, Algeria[1],
Aix Marseille Univ, CNRS, LIS, Marseille, France[3], Aix-Marseille Univ, CNRS, LPL, Aix-en-Provence, France[4],
Institute of Language,Communication and the Brain, France[5]
{jo_silem, h_oufaida}@esi.dz, {maiwenn-annie.fleig, leonor.becerra, philippe.blache}@univ-amu.fr

Abstract

Understanding how language is processed and represented cognitively increasingly involves the use of specialized language models. Yet, because most models are predominantly trained on written text, they struggle to reflect the characteristics of language as it naturally unfolds in spoken interaction. This gap limits their capabilities in capturing features typical of spontaneous speech, such as repetitions, feedback cues, and hesitations. In this work, we introduce linguistically motivated evaluation metrics designed to target these specific spoken-language phenomena. We apply them to analyse outputs from language models fine-tuned on spoken English and French, comparing their behaviour statistically with human dialogue corpora. Our findings highlight the value of these metrics for assessing the degree to which model-generated utterances resemble authentic human conversation.

1 Introduction

The cognitive bases of human language processing are increasingly being explored using large language models across various domains, such as linguistics (Millière, 2024; Piantadosi, 2023) and cognitive neuroscience (Caucheteux et al., 2023; Hosseini et al., 2024). We argue that investigating how humans are capable of producing and understanding language in natural settings, such as conversational interactions, first requires the development of language models specifically adapted to spoken conversation. This presents a unique challenge, distinct from the requirements of classical dialogue systems, as such studies aim not only to replicate human spoken language but also to fully capture all of its dimensions.

Studying the cognitive bases of human language processing can require the estimation of word probabilities for different linguistic phenomena in a conversational context, for instance,to assess the processing difficulty of an event and predict the associated neuro-physiological signals (brain activity, gaze, movements, etc.) (Haller et al., 2024; Smith and Levy, 2013; Frank and Willems, 2017). Language models offer a tool for such estimation, as they are designed to predict word probability distributions. However, in order to use such models in these studies, they must first be adapted to conversational settings, which raises the question of the methodology that can be used to develop such models. Given the complexity of the task, this paper focuses on specific phenomena of spoken language derived from conversation transcriptions without incorporating acoustic or prosodic features. While this is a notable limitation, it represents an essential first step in a largely underexplored area.

The evaluation of large language models typically assesses response quality and relevance within the context. Our objective, however, is different: we seek to determine whether the generated conversations from these models replicate natural spoken language, including its specific phenomena such as pauses, hesitations, and repetitions. To this end, we employ a set of evaluation metrics that provide an analytical perspective. The choice of these metrics is based on different works in descriptive linguistics focusing on spoken language (Candea, 2000; Cook, 1971; Gósy, 2023), which showed that the occurrence of such phenomena during a spontaneous interaction is not random, but can follow certain statistical patterns.

To summarize, this paper introduces a set of linguistic metrics designed specifically to assess the naturalness of the generated conversation. These metrics aim to improve the evaluation of model quality from the perspective of studying the cognitive bases of conversation. To do this, we compare a pre-trained and a fine-tuned language model based on their performance in generating human-like spoken conversational sequences.

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 285–293
Aug 25–27, 2025. ©2025 SIGDIAL

2 Related works

The evaluation of large language models is a difficult task. Previous approaches to language model evaluation mostly relied on lexical-level metrics that compare generated sequences to ground-truth sequences (BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005)) and semantic-level metrics like BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021), which leverage embeddings from pre-trained language models to assess the quality of generated sequences. Given the one-to-many nature of human language and the impressive evolution of large language models, more recent works have shifted toward using more sophisticated approaches to assess models' performance across different domains, including logical and mathematical reasoning (MATH (Hendrycks et al., 2021)), question answering (Mihaylov et al., 2018) and evaluation in multi-turn settings (MultiChallenge (Sirdeshmukh et al., 2025)). Human evaluation has also been widely used to assess the naturalness, coherence, and factuality of LLM outputs (Van der Lee et al., 2021). More recently, foundational models like GPT-4 have been leveraged to evaluate other models in the "LLM-as-judge" paradigm to overcome the time and cost limitations of human evaluation (Chiang and Lee, 2023).

While most of the works focused on in-task evaluation of LLMs, other works have proposed evaluating the linguistic features of these models in off-task settings. In (Reviriego et al., 2023; Martínez et al., 2024), the authors compared the linguistic diversity of LLMs to that of humans, while in (Toro, 2023), the phonological biases of LLMs were studied, showing that these models tend to favor consonants over vowels when identifying words. In another study (Muñoz-Ortiz et al., 2024), a semantic and morphosyntactic evaluation revealed that LLMs still exhibit noticeable differences compared to human-generated text. In our work, we propose an off-task linguistic evaluation of models trained on spoken data by assessing their ability to generate specific phenomena of spoken language.

3 Method

We propose a set of linguistic metrics more specifically adapted to the evaluation of linguistic features of human spoken language in sequences generated by a language model. Among the different phenomena of spoken language, we will pay attention into two particular: disfluencies and feedbacks. Disfluencies refer to any phenomenon that disrupts the smooth, ideal word-to-word flow of speech, such as repetitions, hesitations, and restarts (Corley and Stewart, 2008; Ferreira and Bailey, 2004a). In our study, we will focus on two common disfluencies: repetitions and filled pauses. Repetitions refer to repeated words and phrases that humans produce while talking, while filled pauses (FP) refer to vocalized hesitation pauses that occur in speech which are transcribed with words like '*euh*' in French and '*um*' or '*uh*' in English (Rose, 1998; Candea, 2000).

Drawing inspiration from various studies on linguistics research, which have shown that disfluencies tend to occur at specific grammatical locations or after certain categories of words (Candea, 2000; Rose, 1998) (see Figure S1 in Appendix F, indicating that word repetition is more common in certain categories than others, such as pronouns), we propose two novel metrics for evaluating disfluencies generation by language models based on frequency and on word categories and through comparison to a human reference corpora. First, we propose to evaluate the frequency of occurrences of these phenomena to assess potential over-generation by the model as they are very frequent in human spoken language (Ferreira and Bailey, 2004b). We define the frequency of filled pauses (**Freq-FP**) as the ratio of words that represent a filled pause (words like 'euh' , 'um' and 'uh') in the sequences generated by a model. For repetition (**Freq-Rep**), we took inspiration from Li et al. (2023) and calculated the ratio of repeated words in a sequence using the following score:

$$\textbf{Rep-Token}(seq, n) = \frac{\sum_{k=1}^{n} k * |\texttt{Ngram-Rep}(seq, k)|}{|seq| - n + 1}$$

Where $\texttt{Ngram-Rep}(seq, k)$ is a function that returns the identical, contiguous repeated n-grams of size k in a sequence seq, and n (set to 4 for the study) defines the maximum size of n-grams considered in the detection of the repetition in the sequence. The final **Freq-Rep** was obtained by averaging the **Rep-Token** score for all the sequences generated by the models.

Second, to evaluate the pattern of occurrence of such phenomena, we propose two novel metrics which compare the distribution of the categories of the words (POS tags [1]) preceding filled pauses (**KL-FP**) or the categories of repeated words (**KL-Rep**) in the model-generated sequences to the distribution found in a reference corpus of human spoken language. Our metrics can be formalized as follows:

$$\textbf{KL-Rep} = \exp\left(-\sum_{x\in\mathcal{X}} P_{\text{Rep}}(x)\log\frac{P_{\text{Rep}}(x)}{Q_{\text{Rep}}(x)}\right)$$
$$\textbf{KL-FP} = \exp\left(-\sum_{x\in\mathcal{X}} P_{\text{FP}}(x)\log\frac{P_{\text{FP}}(x)}{Q_{\text{FP}}(x)}\right)$$

Where P_{FP} and P_{Rep} are the distributions of the word categories of repeated words or words preceding a filled pause in the sequences generated by a model, while Q_{FP} and Q_{Rep} are the distributions from a human spoken language reference corpus which is hypothesized to be large enough and diverse to reflect the statistical pattern of the production of disfluencies by a human. The different distributions are compared with KL divergence (Kullback and Leibler, 1951) normalized with a non-linear transformation to obtain a score between 0 and 1 (a higher score is better).

Additionally, we will consider another phenomenon of human spoken language: feedbacks, which refer to the expressions a person produces to convey understanding and interest in what their interlocutor is saying (e.g., '*ok*', '*yeah*') (Boudin et al., 2024). We follow the classification of feedbacks in Boudin et al. (2024) and consider only generic feedbacks, as they are easier to classify given that they consist of a finite set of words (see Appendix C for an example of words that can be found in a generic feedback). For feedbacks, we consider only their frequency defined as the ratio of turns generated by the model that can be classified as feedback.

4 Experiment

To demonstrate our evaluation approaches, we fine-tuned two pre-trained language models : GPT-fr (Radford et al., 2019) and GPT-2 (Simoulin and Crabbé, 2021), on French and English datasets of spoken conversation and compared the models before and after fine-tuning. As our metrics are designed to capture whether sequences follow patterns of spoken language, we expect the scores to be higher after fine-tuning.

[1]The POS tags were determined using the spaCy library (https://spacy.io/), with the `fr_core_news_lg` pipeline for French and `en_core_web_lg` for English.

4.1 Data

For the experiment with the French language, we used the SMYLE corpus (Boudin et al., 2023), an audio-visual and neuro-physiological dataset. SMYLE is a relatively small corpus comprising 30 face-to-face conversations, each conversation involving two French-speaking participants engaged in a storytelling task followed by a free conversation. The data used for the experiments on English conversations comes from the CANDOR corpus (Reece et al., 2023), which includes a total of 1656 conversations held over video chat. For the French reference corpora, we considered CID and ESLO. The CID corpus (Bertrand et al., 2008) is very similar to SMYLE with comparable tasks, i.e. open conversations grounded by storytelling, while the ESLO corpus (Serpollet et al., 2007) is more diverse, including interactive conversations in various contexts. For English, we use the Switchboard Dialog Act Corpus (Stolcke et al., 2000) as a reference, which is an extensive collection of conversations held over telephone. More details on the data can be found in Appendix A.

4.2 Fine-tuning

For the French model, GPT-fr was fine-tuned for 5 epochs using LoRa (Hu et al., 2021) applied to all layers. The English model is a fully fine-tuned large GPT-2 model fine-tuned for 10 epochs. More details on how the models were fine-tuned can be found in Appendix B.

4.3 Sequence Generation

To generate sequences for the evaluation, we used a test subset of the fine-tuning corpora and prompted the models to complete a multi-turn conversation by providing the first seven turns in a conversation as context. See Appendix E for examples of generated turns.

5 Results

The results of the evaluation using the proposed metrics are illustrated in Tables 1 and 2. We additionally report the divergence of word category distributions in the generated sequences, both before and after fine-tuning, compared to the reference and training corpora (**KL-Token**). Overall, we observe different trends across the two languages and metrics. For repetitions, the **Freq-Rep** scores remain relatively similar before and after fine-tuning (less than 3 repeated words per 100 words) and are com-

	Freq-Rep	Freq-FP	Freq-FB
		French	
Before	1.01	4.05	3.04
After	**3.78**	**4.78**	**41.83**
SMYLE	2.34	3.43	20.95
Reference	3.09	3.29	4.73
		English	
Before	**12.0**	0.33	12.77
After	1.21	0.71	**35.53**
CANDOR	1.04	**1.11**	24.01
Reference	1.22	0.25	11.21

Table 1: Results for the repetition frequency (**Freq-Rep**), the frequency of filled pauses (**Freq-FP**), the feedback frequency (**Freq-FB**) for predictions of the models before and after fine-tuning (the results are reported in %). The distribution of the base models -Before- and the fine-tuned models -After- are compared with the training corpus (SMYLE or CANDOR) and the reference corpora (ESLO+CID or Switchboard).

	KL-FP ↑	KL-Rep ↑	KL-Token ↑
		French	
Before/Train	0.336	0.344	0.031
Before/Reference	0.251	0.289	0.028
After/Train	**0.899**	**0.905**	**0.684**
After/Reference	**0.804**	**0.829**	**0.416**
Reference/Train	0.811	0.968	0.338
		English	
Before/Train	0.730	0.859	0.927
Before/Reference	0.454	0.690	0.903
After/Train	**0.964**	**0.936**	**0.977**
After/Reference	**0.670**	**0.835**	**0.972**
Reference/Train	0.745	0.940	0.993

Table 2: Results of KL divergence of the distribution of word categories of predicted words (KL-Token), of word categories preceding filed pauses (KL-FP) and of word categories of repeated words (KL-Rep). Before : Before fine-tuning. After: After fine-tuning.

parable to the scores from the reference and training corpora. The exception is the English model before fine-tuning, which produces sequences with a high frequency of repetition. However, the **KL-Rep** scores significantly improve after fine-tuning, with the most substantial improvement observed in the French model. The improvement after fine-tuning suggests that the models begin to generate repetitions in patterns more aligned with how humans produce repetition in speech. A similar trend is observed for filled pauses. While the overall frequency **Freq-FP** remains close to that of the reference corpora both before and after fine-tuning, the **KL-FP** scores improve significantly after fine-tuning, especially for French, suggesting a better alignment with the distributional patterns of filled pause production in spoken language. Regarding **KL-Token**, the distribution of word categories in the sequences generated by the French model differs considerably from the word categories in the reference spoken corpora and remains very relatively divergent even after fine-tuning. In contrast, the English model shows a closer alignment with the spoken language distributions both before and after fine-tuning.

6 Discussion

The primary contribution of this work is a set of linguistically inspired metrics to evaluate the extent to which language models generate specific phenomena of spoken language. Our experimental results provide insights into the gap between spoken and written language, and how this gap manifests in models trained on the two different forms of language. The frequency-based metrics show that the models can generate various spoken language phenomena, such as repetitions and filled pauses, at rates similar to those found in spoken language corpora. However, when evaluated using divergence-based metrics (KL-Rep and KL-FP), a more nuanced picture emerges: while the frequencies may align, the placement and distributional patterns of these phenomena often do not. For example, although the Freq-Rep scores suggest similar rates of repetition before and after fine-tuning, the low KL-Rep scores before fine-tuning indicate that these repetitions are unnatural and likely the result of text degeneration (Holtzman et al., 2019), where the model loops or repeats phrases unnaturally. This issue is particularly visible in the English model, where frequent repetitions before fine-tuning deviate significantly from natural human repetition patterns. Such observations underscore the utility of divergence-based metrics in revealing qualitative improvements not captured by raw frequency scores.

Overall, the results show that fine-tuning on spoken data improves alignment with human-like generation of repetitions and filled pauses, as reflected in higher KL-Rep and KL-FP scores. However, fine-tuning can also lead to over-generation, particularly of feedback. In the French model, nearly half of the generated turns were identified as feedback, likely due to its high frequency in the training data. This suggests that while fine-tuning helps align models with spoken language, regularization strategies may be necessary to avoid the over-generation of such frequent phenomena. Another key observation is the difference in KL-Token scores between the two languages. In French, the initial KL-Token score before fine-tuning is high, reflecting a mismatch between the word category distributions in the generated sequences and those in spoken corpora—likely due to deeper structural differences between spoken and written French. Fine-tuning reduces this divergence, though it remains more pronounced than in English, where spoken and written language appear to be more similar.

7 Limitations

Our study presents promising results for evaluating how natural, from a linguistic perspective, conversations generated by language models are. However, our evaluation remains relatively basic and additional analysis is needed for a more comprehensive assessment of the fine-tuned models. Future work could explore other phenomena of spoken language, such as turn-taking in the generated conversations or incorporate more advanced syntactic analyses, such as syntactic trees, in comparisons with the reference corpora. Additionally, we acknowledge that the models used in this study are quite outdated compared to recent models such as Llama [2], which may have resulted in the generation of contextually irrelevant sequences or caused the problem of text degeneration observed in the sequences generated by the fine-tuned models. We therefore plan to investigate in future work how fine-tuning larger models on broader datasets may improve the results. Furthermore, while the experiment conducted in this study showed that these metrics can capture hidden aspects of the models such as the generation of degenerate sequences rather than human-like repetitions, further study is needed to assess the robustness of these metrics like exploring how these metrics correlate with human judgment and other evaluation approaches.

Acknowledgment

This work, carried out within the Institute of Convergence ILCB, was supported by grants from France 2030 (ANR-16-CONV-0002). The authors acknowledge that ChatGPT[3] was used to partially check certain phrases in this article, such as spelling and grammatical corrections. However, all coding, writing, and editing were performed by the human authors.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Roxane Bertrand, Philippe Blache, Robert Espesser, Gaëlle Ferré, Christine Meunier, Béatrice Priego-Valverde, and Stéphane Rauzy. 2008. Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. *Revue TAL : traitement automatique des langues*, 49(3).

Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Matthis Houlès, Thierry Legou, Magalie Ochs, and Philippe Blache. 2023. Smyle: A new multimodal resource of talk-in-interaction including neurophysiological signal. In *ICMI'23*.

Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Magalie Ochs, and Philippe Blache. 2024. A multimodal model for predicting feedback position and type during conversation. *Speech Communication*, 159.

Maria Candea. 2000. *Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané. Etude sur un corpus de récits en classe de français.* Ph.D. thesis, Université Paris III.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3):430–441.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Mark Cook. 1971. The incidence of filled pauses in relation to part of speech. *Language and Speech*, 14.

Martin Corley and Oliver W. Stewart. 2008. Hesitation disfluencies in spontaneous speech: The meaning of *um*. 2(4):589–602.

[2]https://ai.meta.com/blog/meta-llama-3/

[3]GPT-4o; http://openai.com

Fernanda Ferreira and Karl G.D. Bailey. 2004a. Disfluencies and human language comprehension. 8(5):231–237.

Fernanda Ferreira and Karl GD Bailey. 2004b. Disfluencies and human language comprehension. *Trends in cognitive sciences*, 8(5):231–237.

Stefan L Frank and Roel M Willems. 2017. Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, 32(9):1192–1203.

Mária Gósy. 2023. Occurrences and durations of filled pauses in relation to words and silent pauses in spontaneous speech. *Languages*, 8(1).

Patrick Haller, Lena S Bolliger, and Lena A Jäger. 2024. On language models' cognitive biases in reading time prediction. In *ICML 2024 Workshop on LLMs and Cognition*. University of Zurich.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Eghbal A Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. 2024. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, 5(1):43–63.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Huayang Li, Tian Lan, Zihao Fu, Deng Cai, Lemao Liu, Nigel Collier, Taro Watanabe, and Yixuan Su. 2023. Repetition in repetition out: Towards understanding neural text degeneration from the data perspective. *Advances in Neural Information Processing Systems*, 36:72888–72903.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Gonzalo Martínez, José Alberto Hernández, Javier Conde, Pedro Reviriego, and Elena Merino-Gómez. 2024. Beware of words: Evaluating the lexical diversity of conversational llms using chatgpt as case study. *ACM Transactions on Intelligent Systems and Technology*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Raphaël Millière. 2024. Language models as models of language. *arXiv preprint arXiv:2408.07144*.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10):265.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL-2002*.

Steven T Piantadosi. 2023. Modern language models refute chomsky's approach to language. *From fieldwork to linguistic theory: A tribute to Dan Everett*, pages 353–414.

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.

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2023. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13):eadf3197.

Pedro Reviriego, Javier Conde, Elena Merino-Gómez, Gonzalo Martínez, and José Alberto Hernández. 2023. Playing with words: Comparing the vocabulary and lexical richness of chatgpt and humans. *arXiv preprint arXiv:2308.07462*.

Ralph Leon Rose. 1998. The communicative value of filled pauses in spontaneous speech. *MA Diss., Univ. of Birmingham*.

Noëlle Serpollet, Gabriel Bergounioux, Annie Chesneau, and Richard Walter. 2007. A large reference corpus for spoken french: Eslo 1 and 2 and its variations. In *Proceedings from Corpus Linguistics Conference Series, University of Birmingham*.

Antoine Simoulin and Benoit Crabbé. 2021. Un modèle transformer génératif pré-entrainé pour le français. In *TALN-2021*. ATALA.

Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. 2025. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. *arXiv preprint arXiv:2501.17399*.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Juan Manuel Toro. 2023. Emergence of a phonological bias in chatgpt. *arXiv preprint arXiv:2305.15929*.

Chris Van der Lee, Albert Gatt, Emiel Van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in neural information processing systems*, 34:27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Data

A.1 French

The SMYLE corpus (Boudin et al., 2023), an audio-visual and neuro-physiological dataset originally collected to study various phenomena related to language production and comprehension and their cognitive processes, is a corpus comprising 30 face-to-face conversations. Each conversation involved two French-speaking participants (mean age = 22.77, SD = 3.29, min = 18, max = 36) engaged in two tasks: a storytelling task (mean duration = 17.49 min, SD = 8.06 min), where one participant had to tell three stories (retelling a video clip shown to the storyteller, describing the plot of a movie or a series, and sharing their favourite holiday story), followed by a free conversation with no specific instructions between the two participants (mean duration = 15.31 min, SD = 3.03 min). For this work, we used the transcriptions of 25 conversations provided in the dataset, which were constructed using a Wav2Vec2 model (Baevski et al., 2020) and manually corrected to add supplementary information to the transcriptions, such as laughter, pronunciation, and disfluencies.

A.2 English

The data used for the experiments investigating English conversations are from the CANDOR corpus (Reece et al., 2023), which includes a total of 1656 conversations held over video chat. The participants are 1456 individuals (52.54% female, 44.17% male, 3.29% other or prefer not to answer, mean age $= 34.32$, std $= 11.42$) which are strangers, who represented a diverse spectrum of gender, educational, ethnic, and generational backgrounds. The corpus provides a total of 850 h of conversations, presenting the audio, video, multiple transcriptions and further features. In this work, only the transcriptions processes by the *audiophile* algorithm are utilized. To eliminate the impact of common initial conversation challenges, like starting a call with "Can you hear me?" — which are unique to non-face-to-face interactions only — the first five exchanges of each conversation were removed from our data subset.

A.3 Data preparation

For this work, we defined a turn as 'the segment of speech produced by a person until interrupted by their interlocutor' to avoid dealing with the overlaps of the IPUs (Interpausal Units). Following this definition, we divided the transcriptions of each conversation into samples of 10 turns, where a turn consists of the consecutive IPU of one person until their interlocutor starts talking (see Appendix D for examples). The turns in each sample were separated by the special token '<p>' and wrapped between BOS and EOS tokens. Filled pauses where represented with their transcriptions, i.e. '*euh*' for French and '*uh*','*um*' or '*uhm*' for English.

For the French corpus, we augmented the dataset with additional samples constructed using random sampling. The corpora were then split into training, validation, and test subsets (80/10/10%). For unbiased evaluation, we ensured that conversations from the same participant were placed in the same subset. Table S1 shows the resulting subsets.

A.4 Reference Corpora

Since no other works use the same evaluation approach as ours (from a linguistic perspective), we will compare our models' results to human spoken language corpora. For French, we consider two reference corpora: CID and ESLO. The CID corpus (Bertrand et al., 2008) is very similar to SMYLE, with comparable tasks i.e., open conversations grounded by storytelling. On the other hand, the ESLO corpus (Serpollet et al., 2007) is more diverse, including interactive conversations in various contexts, from family discussions during meals to interviews and conferences, making it a rich resource for reflecting spoken French in different situations. For English, we use the Switchboard

Language	Subset	Size	#Tokens
French	Training	2800	554400
	Testing	84	14777
	Validation	90	16112
English	Training	91072	7959909
	Testing	11796	1045596
	Validation	11380	1002440

Table S1: The number of samples and tokens in the training, testing, and validation splits of the fine-tuning datasets for both languages.

Dialog Act Corpus (Stolcke et al., 2000) as reference. The Switchboard Dialog Act Corpus is an extensive collection of telephone conversations, where callers pose questions to receivers on a range of topics such as child care, recycling, and news media.

B Model Fine-tuning

B.1 French

For the French model, we used LoRa (Hu et al., 2021) to fine-tune GPT-fr (Simoulin and Crabbé, 2021), a French version of GPT-2. We used the base version of GPT-fr with 1.3B parameters and applied LoRa to all modules across all layers of the model. The model was fine-tuned for 5 epochs using the AdamW optimizer with the following hyperparameter settings: LoRa rank $= 32$, $\alpha = 32$, learning rate $= 2e-4$ with a linear warmup phase of 500 steps, batch size $= 8$, dropout ratio $= 0.05$, and gradient clipping with a norm of 1.

B.2 English

For the English language, the pre-trained language model GPT-2 (Radford et al., 2019) is full parameter fine-tuned on the conversational data. We fine-tuned the large version of GPT-2 with 774M parameters using AdamW optimizer with a learning rate of $5e-5$, a batch size of 1, the maximal block size of 1024 for 10 epochs.

C Generic Feedbacks

Example of words found on a generic feedback :
French: ’*d’accord*’, ’*ouais*’, ’*ouai*’, ’*ok*’, ’*ah*’, ’*mh*’, ’*beh*’, ’*euh*’, ’*heu*’, ’*bah*’
English: ’*mhm*’, ’*uh*’, ’*yeah*’, ’*hmm*’, ’*mm*’, ’*huh*’, ’*um*’

D Examples of training samples

D.1 French Model

<p>ok<p>et du coup euh alexandra c’est ça<p>oui c’est ça<p>du coup tu fais quoi comme étude<p>euh mh<p>enfin si tu fais des études je sais pas ce que tu fais dans la<p>j’ ai fini cet été fin j’ ai arrêté<p>d’accord<p>cet été là je s je commence euh réceptionniste en hôtellerie<p>d’accord

D.2 English Model

<p>Um<p>Mhm<p>So what’s going on?<p>No,<p>Yeah, pumpkin.<p>you see a pumpkin, yep. My son<p>Uh huh.<p>carved that last night. We’re in the middle of Halloween decorating<p>Sure.<p>our house.

E Examples of conversations generated by our models

E.1 French Model

Prompt: <p>salut<p>salut<p>du coup je vais te parler de la première histoire<p>ok<p>et donc il y a un monsieur qui est sur un arbre et qui ramasse des poires

Output Fine-tuned Model: <p>ok<p>et euh il descend de l’ arbre et il les met dans des petits paniers<p>ok<p>donc il descend euh il ramasse les poires il les met dans dans des petits paniers et puis il remonte à l’ arbre et euh il y a un mec qui passe avec une chèvre<p>ok<p>et il passe devant lui il regarde les poires et il voit qu’ il y a trois paniers<p>ok

Output Base Model: mais qui n’en a pas fait un bon
quand il revient en courant il voit le chapeau de son bonhomme mais il le ramasse et il le ramasse en criant ""c’est toi qu’on voulait voir!"" mais en fait il se trouve dans la forêt et c’est son fils qui se fait la belle!
Il demande à son père de lui raconter une belle histoire en lui disant que son père a eu un chapeau et il lui répond qu’il est trop beau pour être le fils de quelqu’un!
Le père le remercie et rentre chez lui avec...

E.2 English Model

Prompt: <p>Yeah, I’m a teacher.<p>Oh,<p>We started back up on monday.<p>you’re a teacher, that’s so interesting. So what are you `teaching?<p>So I`

Output Fine-tuned Model: `<p>Oh,<p> teach high school english, that's my second career, so I'm a high school english teacher and I went to a public school for about a year and a half and then I went back to the public school and now I'm a high school english teacher um<p>okay.<p>because my wife was a music teacher, we got her into the classroom and she did a couple of lessons and she was a very successful music teacher, so I decided to do something else.<p>Oh, so you do the same thing your wife does?`

Output Base Model: `<p>I'm a teacher, I'm a teacher, I'm a teacher, I'm a teacher, I'm a teacher, I'm a teacher, I'm a teacher, I'm a teacher, I'm a teacher, I'm a teacher, I'm a ...`

F Supplementary Figures

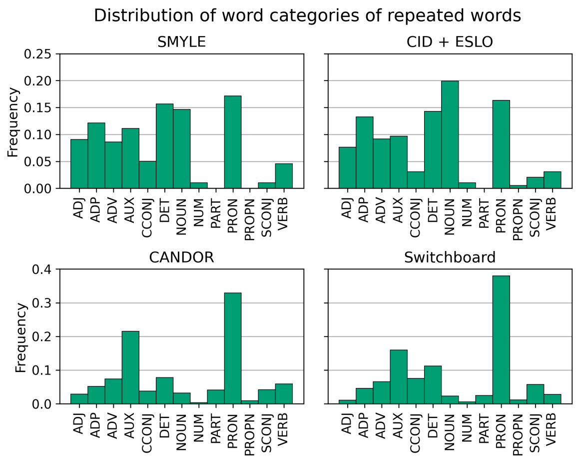

Figure S1: Distribution of the categories of the repeated words in the training and reference corpora for French and English.