Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

A Linguistically-Inspired Approach for the Evaluation of Spoken Language Features in Conversational Models

Oussama Silem[1, 2] Maïwenn Fleig[3,4,5] Philippe Blache[4,5] Houda Oufaida[1]
Leonor Becerra[3,5]
(1) Inria, Paris, France
(2) Ecole nationale Supérieure d'Informatique (ESI ex.INI), Alger, Algérie
(3) Aix Marseille Univ, CNRS, LIS, Marseille, France
(4) Aix-Marseille Univ, CNRS, LPL, Aix-en-Provence, France
(5) Institute of Language,Communication and the Brain, France
jo_silem@esi.dz, maiwenn-annie.fleig@univ-amu.fr, blache@ilcb.fr,
h_oufaida@esi.dz,leonor.becerra@univ-amu.fr

Résumé

Une Approche Linguistique pour l'Évaluation des Caractéristiques du Langage Parlé dans les Modèles Conversationnels

L'étude du traitement du langage et de ses bases cognitives chez l'humain repose de plus en plus sur des modèles de langue adaptés. Cependant, la majorité des modèles existants sont principalement entraînés sur des données écrites, ce qui limite leur pertinence pour l'étude du langage tel qu'il se manifeste dans des contextes naturels, comme lors de conversations spontanées. En effet, ces modèles ne sont pas entraînés pour traiter avec précision les caractéristiques spécifiques du langage parlé, telles que les disfluences et les hésitations. Dans cet article, nous proposons un ensemble de métriques inspirées par la recherche linguistique afin d'évaluer certains phénomènes du langage parlé (feedback, répétition et hésitation) dans des énoncés générés par différents modèles de langue, à travers une comparaison statistique avec des corpus de conversations humaines. Nos résultats, obtenus sur de petits modèles de langue fine-tunés sur des données de conversations parlées en français et en anglais, démontrent le potentiel de ces métriques pour évaluer la similarité des séquences générées avec celles produites par des locuteurs humains.

Abstract

The study of language processing and its cognitive bases increasingly relies on tailored language models. However, most existing language models are trained primarily on written data, limiting their applicability in studying language as it occurs in natural settings, such as in spontaneous conversation, since these models are not trained to accurately handle key features of spoken language like disfluencies and hesitations. In this paper, we propose a set of metrics inspired by linguistic research to evaluate specific phenomena of spoken language (feedback, repetition, and hesitation) in utterances generated by different language models through statistical comparison with corpora of human conversation. Our results, based on small language models fine-tuned on spoken data in French and English, demonstrate the potential of these metrics in assessing the human-likeness of sequences generated by language models.

Mots-clés : Modèles de langue, Conversation spontanée, Évaluation linguistique.

Keywords: Language models, Spontaneous conversation, Linguistic evaluation.

Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM), pages 277–290.

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 conversational agents or 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 features in a conversational context, for instance,to assess the processing difficulty of a linguistic event and predict the associated neuro-physiological signals (brain activity, gaze, movements, etc.) (Haller *et al.*, 2024; Smith & Levy, 2013; Frank & Willems, 2017; Weissbart *et al.*, 2020). 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 spoken conversational settings. This raises the question of which methodology can be used to develop such models, as spoken language differs from the written language on which most recent language models are trained (Mousavi *et al.*, 2024; Kim *et al.*, 2021).

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.

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 & 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 (Minaee *et al.*, 2024), 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), coding (ARCADE (Yin *et al.*, 2023)), 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 & 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 the context of spoken dialogue, several studies have explored the potential of using pretrained language models for the development of spoken dialogue systems. Early work focused on leveraging pretrained language models to improve language understanding tasks (Kim *et al.*, 2021; Yoon *et al.*, 2023). More recent works, such as (Mousavi *et al.*, 2024), examined the robustness of large language models on spoken language, finding that these models are not sufficiently robust to spoken noise. In (Louradour *et al.*, 2024), Louradour *et al.* proposed a set of models for spoken language by continuing the training of large language models on a large dataset of spoken conversation transcripts. While our paper does not aim to build such models, we propose an off-task linguistic evaluation of models trained on spoken data by assessing their ability to generate specific phenomena characteristic 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 & Stewart, 2008; Ferreira & Bailey, 2004). 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 difluencies 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 & Bailey, 2004). 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 corpus. 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. The different distributions are compared with KL divergence (Kullback & Leibler, 1951) normalized with a non-linear transformation to obtain a score between 0 and 1 (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 the evaluation of 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 different pre-trained language models, GPT-fr (Simoulin & Crabbé, 2021) and GPT-2 (Radford *et al.*, 2019), on French and English datasets of spoken conversation and compared the models before and after finetuning. As our metrics are designed to capture whether sequences follow patterns of spoken language, we expect the scores to be

1. The POS tags were determined using the spaCy library [2], with the `fr_core_news_lg` pipeline for French and `en_core_web_lg` for English.

higher after finetuning. Details on the data as well as the fine-tuned models can be found in Appendix A and B. To generate sequences for the evaluation, we used a test subset of the finetuning 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. For reference corpora, we used the ESLO and CID corpus (Serpollet *et al.*, 2007; Bertrand *et al.*, 2008) for French and the Switchboard Dialog Act Corpus (Stolcke *et al.*, 2000) for English.

	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 French and English models before and after finetuning (the results are reported in %). The distribution of the base models -Before- and the finetuned models -After- are compared with the training corpus (SMYLE or CANDOR) and the reference corpora (ESLO+CID or Switchboard).

5 Results

The results of the evaluation using the proposed metrics are reported in Tables 1 and 2. We additionally report the divergence of word category distributions in the generated sequences, both before and after finetuning, 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 finetuning (less than 3 repeated words per 100 words) and are comparable to the scores from the reference and training corpora. The exception is the English model before finetuning, which produces sequences with a high frequency of repetition. However, the **KL-Rep** scores improve after finetuning, with the most substantial improvement observed in the French model. The improvement after finetuning 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 finetuning, the **KL-FP** scores improve significantly after finetuning, 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 finetuning. In contrast, the English model shows a closer alignment with the spoken language distributions both before and after finetuning. For feedbacks, the results show a high frequency of feedback turns in both languages. Interestingly, the different corpora display significant variation in feedback frequency, particularly in French, likely due to differences in the nature of the tasks in each corpus.

	KL-PR	**KL-Rep**	**KL-Token**	**KL-PR**	**KL-Rep**	**KL-Token**
Model		**French**			**English**	
Before/Train	0.336	0.344	0.031	0.730	0.859	0.927
Before/Reference	0.251	0.289	0.028	0.454	0.690	0.903
After/Train	0.899	0.905	0.684	0.964	0.936	0.977
After/Reference	0.804	0.829	0.416	0.670	0.835	0.972
Reference/Train	0.811	0.968	0.338	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 finetuning. After : After finetuning.

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 finetuning, the low KL-Rep scores before finetuning 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 finetuning 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 finetuning 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, finetuning can also lead to over-generation, particularly of feedback. In the French model, nearly half of the generated turns were identified as feedbacks, likely due to its high frequency in the training data. This suggests that while finetuning 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 finetuning 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. Finetuning 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 like Llama [3], which may have resulted in the generation of contextually irrelevant sequences or caused the problem of text degeneration. We therefore plan to investigate in future work how finetuning larger models on larger datasets may improve the results.

Acknowledgment

The authors acknowledge that ChatGPT [4] 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

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, **33**, 12449–12460.

BANERJEE S. & LAVIE A. (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*, p. 65–72.

BERTRAND R., BLACHE P., ESPESSER R., FERRÉ G., MEUNIER C., PRIEGO-VALVERDE B. & RAUZY S. (2008). Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. *Revue TAL : traitement automatique des langues*, **49**(3).

BOUDIN A., BERTRAND R., RAUZY S., HOULÈS M., LEGOU T., OCHS M. & BLACHE P. (2023). Smyle : A new multimodal resource of talk-in-interaction including neuro-physiological signal. In *ICMI'23*.

BOUDIN A., BERTRAND R., RAUZY S., OCHS M. & BLACHE P. (2024). A multimodal model for predicting feedback position and type during conversation. *Speech Communication*, **159**.

CANDEA M. (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.* Thèse de doctorat, Université Paris III.

CAUCHETEUX C., GRAMFORT A. & KING J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, **7**(3), 430–441.

CHIANG C.-H. & LEE H.-Y. (2023). Can large language models be an alternative to human evaluations ? In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st*

3. https ://ai.meta.com/blog/meta-llama-3/
4. GPT-4o ; http ://openai.com

Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), p. 15607–15631, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.870.

COOK M. (1971). The incidence of filled pauses in relation to part of speech. *Language and Speech*, **14**.

CORLEY M. & STEWART O. W. (2008). Hesitation disfluencies in spontaneous speech : The meaning of um. *Language and Linguistics Compass*, **2**(4), 589–602.

FERREIRA F. & BAILEY K. G. (2004). Disfluencies and human language comprehension. *Trends in cognitive sciences*, **8**(5), 231–237.

FRANK S. L. & WILLEMS R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, **32**(9), 1192–1203.

GÓSY M. (2023). Occurrences and durations of filled pauses in relation to words and silent pauses in spontaneous speech. *Languages*, **8**(1), 79.

HALLER P., BOLLIGER L. S. & JÄGER L. A. (2024). On language models' cognitive biases in reading time prediction. In *ICML 2024 Workshop on LLMs and Cognition* : University of Zurich.

HENDRYCKS D., BURNS C., KADAVATH S., ARORA A., BASART S., TANG E., SONG D. & STEINHARDT J. (2021). Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv :2103.03874*.

HOLTZMAN A., BUYS J., DU L., FORBES M. & CHOI Y. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv :1904.09751*.

HOSSEINI E. A., SCHRIMPF M., ZHANG Y., BOWMAN S., ZASLAVSKY N. & FEDORENKO E. (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.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L., CHEN W. *et al.* (2022). Lora : Low-rank adaptation of large language models. *ICLR*, **1**(2), 3.

KIM S., LIU Y., JIN D., PAPANGELIS A., GOPALAKRISHNAN K., HEDAYATNIA B. & HAKKANI-TÜR D. (2021). "how robust r u ?" : Evaluating task-oriented dialogue systems on spoken conversations. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, p. 1147–1154. DOI : 10.1109/ASRU51503.2021.9688274.

KULLBACK S. & LEIBLER R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, **22**(1), 79–86.

LI H., LAN T., FU Z., CAI D., LIU L., COLLIER N., WATANABE T. & SU Y. (2023). Repetition in repetition out : Towards understanding neural text degeneration from the data perspective. *Advances in Neural Information Processing Systems*, **36**, 72888–72903.

LIN C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. In *Text summarization branches out*, p. 74–81.

LOURADOUR J., HUNTER J., HARRANDO I., SHANG G., RENNARD V. & LORRÉ J.-P. (2024). Claire : Large language models for spontaneous French dialogue. In *TALN-2024*.

MARTÍNEZ G., HERNÁNDEZ J. A., CONDE J., REVIRIEGO P. & MERINO-GÓMEZ E. (2024). Beware of words : Evaluating the lexical diversity of conversational llms using chatgpt as case study. *ACM Transactions on Intelligent Systems and Technology*.

MIHAYLOV T., CLARK P., KHOT T. & SABHARWAL A. (2018). Can a suit of armor conduct electricity ? a new dataset for open book question answering. *arXiv preprint arXiv :1809.02789*.

MILLIÈRE R. (2024). Language models as models of language. *arXiv preprint arXiv :2408.07144*.

MINAEE S., MIKOLOV T., NIKZAD N., CHENAGHLU M., SOCHER R., AMATRIAIN X. & GAO J. (2024). Large language models : A survey, 2024. *arXiv preprint arXiv :2402.06196*.

MOUSAVI S. M., ROCCABRUNA G., ALGHISI S., RIZZOLI M., RAVANELLI M., RICCARDI G. *et al.* (2024). Are llms robust for spoken dialogues ? In *Proceedings of the 14th International Workshop on Spoken Dialogue Systems Technology* : 14th International Workshop on Spoken Dialogue Systems Technology.

MUÑOZ-ORTIZ A., GÓMEZ-RODRÍGUEZ C. & VILARES D. (2024). Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, **57**(10), 265.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *ACL-2002*.

PIANTADOSI S. T. (2023). Modern language models refute chomsky's approach to language. *From fieldwork to linguistic theory : A tribute to Dan Everett*, p. 353–414.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D., SUTSKEVER I. *et al.* (2019). Language models are unsupervised multitask learners. *OpenAI blog*, **1**(8), 9.

REECE A., COONEY G., BULL P., CHUNG C., DAWSON B., FITZPATRICK C., GLAZER T., KNOX D., LIEBSCHER A. & MARIN S. (2023). The candor corpus : Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, **9**(13), eadf3197. DOI : 10.1126/sciadv.adf3197.

REVIRIEGO P., CONDE J., MERINO-GÓMEZ E., MARTÍNEZ G. & HERNÁNDEZ J. A. (2023). Playing with words : Comparing the vocabulary and lexical richness of chatgpt and humans. *arXiv preprint arXiv :2308.07462*.

ROSE R. L. (1998). The communicative value of filled pauses in spontaneous speech. *MA Diss., Univ. of Birmingham*.

SERPOLLET N., BERGOUNIOUX G., CHESNEAU A. & WALTER R. (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*.

SIMOULIN A. & CRABBÉ B. (2021). Un modèle transformer génératif pré-entrainé pour le français. In P. DENIS, N. GRABAR, A. FRAISSE, R. CARDON, B. JACQUEMIN, E. KERGOSIEN & A. BALVET, Éds., *TALN-2021* : ATALA.

SIRDESHMUKH V., DESHPANDE K., MOLS J., JIN L., CARDONA E.-Y., LEE D., KRITZ J., PRIMACK W., YUE S. & XING C. (2025). Multichallenge : A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. *arXiv preprint arXiv :2501.17399*.

SMITH N. J. & LEVY R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, **128**(3), 302–319.

STOLCKE A., RIES K., COCCARO N., SHRIBERG E., BATES R., JURAFSKY D., TAYLOR P., MARTIN R., VAN ESS-DYKEMA C. & METEER M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, **26**(3), 339–374.

TORO J. M. (2023). Emergence of a phonological bias in chatgpt. *arXiv preprint arXiv :2305.15929*.

VAN DER LEE C., GATT A., VAN MILTENBURG E. & KRAHMER E. (2021). Human evaluation of automatically generated text : Current trends and best practice guidelines. *Computer Speech & Language*, **67**, 101151.

WEISSBART H., KANDYLAKI K. & REICHENBACH T. (2020). Cortical tracking of surprisal during continuous speech comprehension. pg - 155-166. *Journal of cognitive neuroscience*, **32**(1), 155–166.

YIN P., LI W.-D., XIAO K., RAO A. K., WEN Y., SHI K., HOWLAND J., BAILEY P., CATASTA M., MICHALEWSKI H. *et al.* (2023). Natural language to code generation in interactive data science notebooks. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

YOON J., HWANG S., KIM K.-E., RAN H. & BANG J.-U. (2023). Adapting text-based dialogue state tracker for spoken dialogues. In *24th SIGdial Workshop on DSTC11 : The Eleventh Dialog System Technology Challenge*, p. 81–88 : Association for Computational Linguistics.

YUAN W., NEUBIG G. & LIU P. (2021). Bartscore : Evaluating generated text as text generation. *Advances in neural information processing systems*, **34**, 27263–27277.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). Bertscore : Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Data

A.1 French

For the experiment with the French language, we used 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. SMYLE is a relatively small 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 favorite 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 is 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

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 finetuning datasets for both languages.

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 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 Training

B.1 French

For the French language, we used LoRa (Hu *et al.*, 2022) to fine-tune GPT-fr (Simoulin & 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*', '*ouaï*', '*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 Finetuned 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 Finetuned 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 ...`

F Supplementary Figures

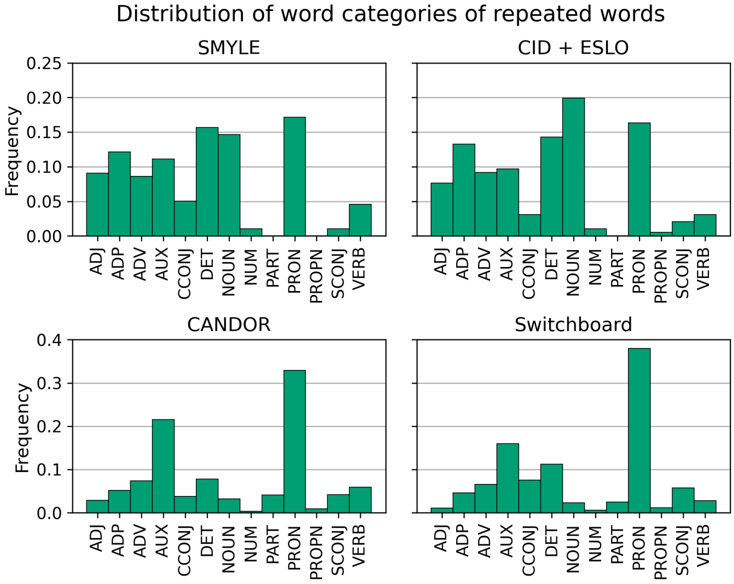

FIGURE S1 – Distribution of the categories of the repeated words in the training and reference corpora for French and English.