Análise de comportamento do modelo NotebookLM em transcrição automática

Brenda R. F. de Oliveira[1], **Oto Araújo Vale**[1]

[1]Departamento de Letras – Universidade Federal de São Carlos (UFSCar)
13565-905 – São Carlos – SP – Brasil

{brendafernandes@estudante.ufscar.br, otovale@ufscar.br}

Abstract. *This research analyzes the behavior of the NotebookLM model in automatic speech transcription, using a corpus of 55 economic interviews from the Roda Viva TV program. The study is based on the hypothesis that the model's textual registry choices reveal both its capabilities and limitations regarding spontaneous speech. By identifying and organizing behavioral patterns, a typology representing the model's behavioral particularities was constructed to form an annotation guideline. Results obtained so far demonstrate that the model alternates between sensitivity to speech disfluencies and limitations in textual organization.*

Resumo. *Esta pesquisa analisa o comportamento do modelo NotebookLM na transcrição automática da fala, tomando como corpus a transcrição de 55 entrevistas de economia do programa Roda Viva. O estudo parte da hipótese de que as escolhas de registro textual do modelo revelam tanto suas capacidades quanto suas limitações diante da fala espontânea. Com a identificação e organização de padrões de comportamento, construiu-se uma tipologia que representa as particularidades de comportamento do modelo para compor uma diretriz de anotação. Os resultados obtidos até o momento demonstram que o modelo alterna entre sensibilidade às disfluências da fala e limitações na organização textual.*

1. Quadro teórico e objetivo de pesquisa

A presente pesquisa tem como objetivo analisar sistematicamente o comportamento do modelo NotebookLM[1] na tarefa de transcrição automática da fala, tomando como base um conjunto de entrevistas de economia do programa Roda Viva da TV Cultura. Nesse sentido, o interesse central reside em compreender as tendências comportamentais e limitações do sistema computacional, ao identificar padrões recorrentes em suas produções textuais, avaliar escolhas feitas pelo modelo e a forma como ele lida com a natureza complexa da fala. Tomou-se como premissa de que, além de ser caracterizada por uma natureza dinâmica e fragmentária [Clark and Fox Tree 2002, Galembeck 1999], a fala espontânea apresenta caráter quase simultâneo na relação entre seu planejamento e a sua realização [Duarte et al. 2014]. Nesse contexto, a fala não é compreendida como produto acabado, mas como processo cuja materialidade permite que existam marcas de planejamento em sua superfície e estrutura, conhecidas como disfluências [Madureira 2026].

[1]notebooklm.google.com

A relevância crescente das tecnologias baseadas em fala, como assistentes de voz, sistemas de reconhecimento automático de fala (ASR) e plataformas de diálogo, demonstra que a transcrição automática ocupa papel fundamental. Isso se deve a qualidade das interações entre humano e máquina que depende, em grande medida, da precisão com que os sinais sonoros são convertidos em linguagem escrita. No entanto, a fala espontânea representa um desafio particular para Grandes Modelos de Linguagem Multimodais (MLLM), pois é mais complexa de processar e transcrever com precisão, [Casanova et al. 2026], além de carecer de estrutura clara, ser fragmentada e apresentar disfluências. Com isso, parte-se do pressuposto de que o comportamento do modelo, materializado nas escolhas de registro textual, pode revelar suas limitações técnicas, bem como sua capacidade de lidar com a complexidade constitutiva da fala em interação.

2. Metodologia

A metodologia utiliza abordagem qualitativa e descritiva, com o apoio da quantificação das ocorrências observadas. Nesse sentido, partiu-se de um *corpus* composto pela transcrição de 55 entrevistas de economia do programa Roda Viva que foram realizadas entre 2016 e 2026 e estão disponíveis no canal oficial do programa no YouTube[2]. Tratam-se de produções de longa duração em língua portuguesa brasileira e de livre acesso ao público. Pesquisas recentes utilizam *corpora* ideais para analisar a fala espontânea autêntica, como o NURC-SP [Lima et al. 2024] e o CORAA ASR [Gris et al. 2022], e avaliar modelos como Wav2Vec2 e Whisper por meio de métricas padrão como a *Word Error Rate* (WER). Diferentemente desses *corpora*, reconhece-se que as entrevistas, relacionadas ao *corpus* do presente trabalho, não representam integralmente a fala espontânea, pois o discurso tende a apresentar formalidade e monitoramento acentuados.

Entretanto, tais contextos de interação preservam características da fala espontânea como hesitações, reformulações, repetições, assim como a simultaneidade entre planejamento e produção. Isso aproxima tais interações de circunstâncias de fala em uso e as diferem de situações de oralização da escrita[3], tornando-as adequadas para a análise deste estudo. Além disso, a métrica WER avalia o desempenho lexical de sistemas de ASR com a normalização de transcrições e cálculo. O presente trabalho, no entanto, propõe uma investigação para além da avaliação mecânica, ao preservar as transcrições e incluir abordagens qualitativas, considerando aspectos[4] como a segmentação, a capitalização, a pontuação, além do registro de fenômenos da fala espontânea.

O arquivo de áudio de cada entrevista foi importado para a plataforma NotebookLM que realizou automaticamente a conversão dos sinais sonoros em texto escrito sem oferecer a opção para incluir prompt. Com isso, não foi fornecido nenhuma instrução adicional, permitindo preservar o funcionamento original do modelo. As transcrições resultantes foram organizadas em um editor de texto para análise. Na etapa de identificação manual de padrões de comportamento do modelo, as ocorrências foram sinalizadas no

[2] `www.youtube.com/rodaviva`

[3] A leitura em voz alta e a realização de discursos roteirizados são alguns exemplos de oralização da escrita.

[4] Embora não sejam contemplados diretamente pela métrica WER, acredita-se que tais fatores também influenciam a qualidade e a precisão da transcrição tanto no âmbito lexical quanto na interpretação da informação.

próprio corpo do texto e classificadas segundo sua natureza. Nesse processo, foi realizada a comparação manual das transcrições com o auxílio do vídeo das respectivas entrevistas, permitindo verificar a equivalência entre o conteúdo verbal e o texto escrito. Em casos de incerteza, os trechos correspondentes foram reproduzidos novamente, a fim de confirmar a classificação atribuída. As ocorrências anotadas foram registradas em planilhas por meio de ferramentas automatizadas, para a contagem das frequências e a verificação sistemática dos padrões identificados. Essa organização permitiu desenvolver uma tipologia que representa as particularidades de comportamento do modelo para compor uma diretriz que orienta a anotação das demais transcrições presentes no *corpus*.

3. Resultados e discussão

Até o momento, foram identificadas dez categorias, distribuídas em três grandes grupos, sendo a primeira voltada para o registro de marcas de disfluência e planejamento da fala; o segundo para os desvios de normatização e escrita; e o terceiro voltado para inadequações estruturais e de conteúdo. No primeiro grupo, foram observadas categorias como (i) Registro de repetição, (ii) Registro de hesitação e (iii) Registro de reformulação da fala. No segundo, incluem-se (i) Desvio ortográfico, cujas subcategorias são (i.a) Palavra e (i.b) Número; (ii) Desvio de pontuação[5] e (iii) Desvio de capitalização. Já o terceiro grupo abrange (i) Inadequação de segmentação[6], (ii) Omissão[7], (iii) Inserção e Substituição lexical, sendo esta última subdividida em (i.a) Por homofonia e (i.b) Por alucinação[8].

Tendo isso em vista, entende-se que as categorias presentes tanto no primeiro quanto no segundo grupo revelam as limitações técnicas e as escolhas do modelo na tarefa de transcrever a fala. A análise inicial indica que ele demonstrou dificuldade em aplicar as normas da escrita a um material da fala que possui especificidades próprias. Já as categorias presentes no primeiro grupo refletem a sensibilidade do modelo às disfluências naturais da fala. Isso pode ser observado no registro de hesitação em que o modelo reconhece o fenômeno e representa, em diversos casos, como "eh" em vez de "é". A frequência que as categorias apresentam são variadas e podem ser observadas com os dados da transcrição da entrevista de Ricardo Alban[9] como exemplo presente na tabela 1.

Nessa entrevista, os dados mostram que os desvios ortográficos e registros de repetição aparecem com maior incidência, seguidos por registros de hesitação e desvios de capitalização. Isso sugere que o modelo alterna entre sensibilidade às disfluências da fala e falhas na organização textual. Nessa análise inicial, o modelo também demonstrou falhar em capturar certos trechos da fala humana como, por exemplo, a fala "O ministro Fernando Haddad [não dá os detalhes da medida], mas já adiantou que boa parte delas pre-

[5]Nessa categoria, considera-se as ocorrências com uso de sinais gráficos em desacordo com as normas gramaticais. Esse desvio pode ocorrer pelo registro de determinada pontuação em situação que não permite sua inclusão ou pela ausência de pontuação.

[6]Nessa categoria, considera-se as ocorrências em que palavra ou trecho é escrito com separação em desacordo com a norma-padrão, resultando em uma ocorrência de hipossegmentação ou hipersegmentação.

[7]A omissão ocorre quando parte do discurso gravado não é registrada na transcrição. Ocorrências caracterizadas pela ausência de letras dentro de uma palavra não fazem parte dessa categoria, mas da "Desvio ortográfico".

[8]Considera-se como alucinação, quando o modelo substitui palavras, presentes no áudio original, por outras que não foram ditas.

[9]Presidente da Confederação Nacional da Indústria (CNI), entrevista realizada em 04/08/2025 - `https://www.youtube.com/watch?v=EjOvy76WA2M`

Tabela 1. Frequência de ocorrência por categoria

Natureza da categoria	Categorias	Frequência
Marcas de disfluência e planejamento da fala	Registro de repetição	121
	Registro de hesitação	93
	Registro de reformulação da fala	51
Desvios de normatização e escrita	Desvio ortográfico	169
	Palavra	167
	Número	2
	Desvio de pontuação	41
	Desvio de capitalização	82
Inadequações estruturais e de conteúdo	Inadequação de segmentação	27
	Omissão	63
	Inserção	14
	Substituição lexical	41
	Por homofonia	36
	Por alucinação	5

cisaria passar pelo Congresso Nacional de um jeito ou de outro, né?", em que o fragmento delimitado por colchetes estava presente na fala gravada, mas foi omitido na transcrição pelo modelo. Nesse sentido, o NotebookLM não apenas reconheceu certos aspectos da oralidade, mas também impôs uma reorganização linguística sobre a fala que, nesse caso, comprometeu a interpretação do que foi dito.

4. Conclusão

Este estudo une Linguística e Processamento de Língua Natural. Assim, espera-se não apenas descrever padrões de comportamento do sistema, mas também contribuir para avaliações mais refinadas de tecnologias de transcrição automática no português brasileiro. As limitações manifestadas pelo NotebookLM demonstram que ele, enquanto um MLLM, apresentou dificuldade em aplicar automaticamente as regras formais da escrita na transcrição da fala humana. Por outro lado, diferente de sistemas de conversão mecânica de sinais sonoros em texto, o modelo NotebookLM revelou sensibilidade ao planejamento da fala, com a inferência contextual necessária para reconhecer e distinguir tanto hesitações como "eh", de formas verbais como "é", quanto às demais disfluências, reforçando habilidades de reorganização linguística.

Considerando que esta pesquisa ainda está em desenvolvimento, cabe ressaltar que os resultados apresentados possuem caráter preliminar, pois estão voltados apenas para a etapa inicial da investigação e referem-se exclusivamente ao NotebookLM. Pretende-se realizar posteriormente uma comparação com os resultados obtidos pelo modelo Whisper para verificar se os padrões observados são específicos do NotebookLM ou compartilhados por outros sistemas. Nesse sentido, as discussões realizadas não buscam propor conclusões definitivas sobre o conjunto de entrevistas, mas ilustrar os padrões observados até o momento. Portanto, eventuais generalizações poderão ser realizadas somente após a conclusão da análise integral do *corpus* e da etapa comparativa.

Referências

Casanova, E., Santos, V. G., Svartman, F. R. F., Leite, M. Q., Candido Junior, A., Marcacini, R. M., Rezende, S. O., and Aluísio, S. M. (2026). Recursos para o processamento de fala. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, volume 1, book chapter 3. BPLN, 4 edition.

Clark, H. H. and Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Duarte, T. P. P. J., Storto, L. J., and Durante, D. (2014). Características da língua falada e da língua escrita: análise de entrevistas com vik muniz.

Galembeck, P. d. T. (1999). Metodologia de pesquisa em português falado. In Rodrigues, A. C. d. S., Alves, I. M., and Goldstein, N. S., editors, *I Seminário de Filologia e Língua Portuguesa*, pages 109–119. Humanitas Publicações FFLCH/USP, São Paulo.

Gris, L., Junior, A. C., Santos, V., Dias, B., Leite, M., Svartman, F., and Aluísio, S. (2022). Bringing nurc/sp to digital life: the role of open-source automatic speech recognition models. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 330–341, Porto Alegre, RS, Brasil. SBC.

Lima, R., Leal, S., Junior, A. C., and Aluísio, S. (2024). A large dataset of spontaneous speech with the accent spoken in são paulo for automatic speech recognition evaluation. In *Anais da XXXIV Brazilian Conference on Intelligent Systems*, pages 33–47, Porto Alegre, RS, Brasil. SBC.

Madureira, B. (2026). Diálogo e interatividade. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, volume 1, book chapter 15. BPLN, 4 edition.