Corpus design

1. Perguntas de Pesquisa

1. Que tipos de discursos sobre a mudança climática são mobilizados e circulam em diferentes registros na internet, por diferentes atores sociais, no período contemporâneo?
2. Em que medida os discursos sobre a mudança climática, identificados em dimensões discursivas nos registros analisados, distinguem a produção textual humana de textos correspondentes gerados por modelos de linguagem de grande porte (LLMs)?
3. Que consonâncias (alinhamentos, reforços semânticos ou estabilização de sentidos) e que tensões (divergências, deslocamentos discursivos ou resistências) emergem da comparação entre os discursos produzidos por atores humanos e aqueles gerados por LLMs?

2. Princípios Gerais do Corpus Design

O corpus é estruturado a partir de **estratos híbridos**, definidos pela combinação entre:

- **ator social** (quem produz o texto) e
- **registro discursivo** (que tipo de texto é produzido).

Essa decisão decorre da orientação de que não é conceitualmente válido agrupar textos apenas por ator social quando pertencem a registros distintos (por exemplo, atas e relatórios no caso de instituições governamentais).

Cada estrato contém **200 unidades de amostragem (sampling units)**, o que assegura:

- equilíbrio entre estratos;
- condições mínimas para análises fatoriais estáveis;
- robustez para análises lexicais comparativas;
- paridade entre textos humanos e textos gerados por LLMs.

O corpus final é composto por aproximadamente **2.000 textos**, sendo:

- 1.000 textos humanos;
- 1.000 textos gerados por LLMs.

Esse número decorre diretamente do desenho de balanço entre estratos e da exigência de estabilidade analítica, e não de critérios de custo ou conveniência.

“O tamanho total do corpus resulta do desenho metodológico, do número de estratos e do balanceamento entre eles, aliado à exigência de paridade entre textos humanos e textos gerados por modelos de linguagem e à necessidade de volume suficiente para análises lexicais e fatoriais estáveis.”

3. Estrutura do Subcorpus Humano (1.000 sampling units)

Cada estrato corresponde a uma combinação específica de ator social e registro discursivo:

1. **Governo – Ata (Decisões COP)**
    - 200 sampling units
    - Unidades extraídas como **segmentos textuais autênticos** das atas.
2. **Governo – Relatório (Relatórios do IPCC)**
    - 200 sampling units
    - Unidades extraídas como **segmentos textuais autênticos** dos relatórios.
3. **Imprensa – Reportagem (Now)**
    - 200 textos.
4. **ONGs – Story (WWF, WRI, Greenpeace)**
    - 200 textos (66, 67, 67)
    - Distribuição equilibrada entre as três ONGs.
5. **Academia – Abstract (AntCorGen)**
    - 200 textos.
6. **Rede social – Postagem (Gettr)**
    - 200 textos.

(**Neste estudo, o discurso de redes sociais é operacionalizado por meio do estrato *Rede social – Postagem textual*. O Gettr é adotado para servir como fonte empírica para a construção do strata, configurando uma decisão de natureza operacional, orientada por critérios de viabilidade e acessibilidade textual, e não por uma pretensão de representatividade das plataformas de redes sociais como um todo.**)

4. Definição da Sampling Unit

A unidade de amostragem não corresponde ao texto integral, mas a um **segmento textual autêntico**.

Essa decisão atende simultaneamente a duas exigências metodológicas:

1. **Exigência técnica**
   Textos longos (como atas completas e relatórios extensos) não podem ser integralmente reproduzidos por LLMs.
2. **Exigência analítica**
   A análise discursiva e lexical exige unidades comparáveis entre textos humanos e textos gerados.

Procedimento de segmentação

- Identificar, em cada documento longo, seções relevantes e informativas (por exemplo, introduções, seções de desafios, sínteses);
- Extrair múltiplos segmentos por documento até atingir 200 sampling units por strata;
- Registrar manualmente o início e o fim de cada segmento (begin / end) para posterior extração automática.

Essa estratégia permite aproveitar documentos escassos (por exemplo, os relatórios do IPCC) por meio de segmentação, em vez de descartar material.

5. Subcorpus LLMs (1.000 textos)

4 LLMs (que não precisa baixar na máquina, a coleta é online): ChatGPT, Gemini, Grok e Claude (250 de cada):
ChatGPT – 250

Gemini- 250

Grok – 250

Claude – 250

1000 textos no total

Os textos gerados por LLMs espelham exatamente os mesmos estratos e sampling units do subcorpus humano.

Cada sampling unit humana gera um texto correspondente por LLM.

Todos os textos são produzidos por **um único tipo de prompt**.

Não há duplicação do corpus por tipo de prompt.

6. Desenho dos Prompts

6.1 Tipo de prompt adotado

Utiliza-se exclusivamente um **prompt detalhado**, com reconstrução da **cena** (contexto de produção textual).

Não se utiliza prompt default.

Não vamos combinar diferentes tipos de prompt.

6.2 Estrutura do prompt

a) System prompt
Uma única frase que define:

- ocupação;
- papel institucional;
- função social do agente produtor do texto.

Não se introduzem traços ideológicos nem características de personalidade.

b) User prompt
Especifica:

- ator social;
- registro discursivo;
- seção textual;
- extensão aproximada;
- informações extraídas da sampling unit humana.

c) Resumo prévio (≤ 100 palavras)

- Gerado a partir da sampling unit humana;
- Formulado nas palavras do próprio LLM;
- Não idêntico ao texto original.

Esse resumo funciona como **input informacional** para o prompt.

6.3 Critério central

O desenho do prompt controla a **informação**, não o **discurso**.

Assim, toda variação posterior entre textos humanos e textos gerados é interpretada como variação discursiva produzida pelo LLM.

7. Reconstrução da Cena (Contexto de Produção)

A geração textual é orientada pela reconstrução da **cena**, que inclui:

- ator social;
- papel institucional;
- registro;
- finalidade comunicativa;
- situação de produção.

Essa reconstrução busca fornecer a maior quantidade possível de informação contextual para aproximar a produção do LLM das condições reais de produção dos textos humanos.

COMPLEMENTOS ROGÉRIO:

“A distinção entre System Prompt e User Prompt ocorre no nível interno do sistema do LLM. O System Prompt especifica ao LLM qual papel ou qual pessoa ele deve assumir (por exemplo: 'Você é um professor universitário da área de Linguística (papel), pardo e de classe média-alta (pessoa).'). Ele especifica quem o LLM deve ser para realizar a tarefa.

O User Prompt especifica a tarefa que o LLM deverá realizar (por exemplo: 'Prepare uma aula sobre Análise Multidimensional Tradicional para os novos estudantes do curso ‘tarefa’).

Normalmente, o System Prompt é oculto ao usuário pelo desenvolvedor do Agente de IA como forma de personalizar a experiência do usuário. No exemplo, o usuário se sente como se estivesse tratando com um mentor especialista em Linguística.

O User Prompt é a caixa de interação que o usuário usa para interagir com o LLM, daí o nome. Contudo, o usuário pode usar o User Prompt para redefinir o System Prompt, combinando os dois. Por exemplo: 'Você é um professor universitário da área de Linguística (papel), pardo e de classe média-alta (pessoa). Prepare uma aula sobre Análise Multidimensional Tradicional para os novos estudantes do curso (tarefa).'

EX:

System Prompt: 'You are a senior member of the Working Group 1, Scientific Assessment of Climate Change, who is editing an Intergovernmental Panel on Climate Change (IPCC) report.

User Prompt (+ 'Cena Prompt'): Write the Introduction of the IPCC report considering the following summary. Do not invent information. Do not assign

titles or subtitles. Do not acknowledge this prompt, just provide the response straightaway.

<summary>

O sumário especifica a cena. Poderia ser descrição de estilo a ser seguido, etc.”

8. Construção da Lista de Palavras para Coleta de Textos

“O Mark Davis, por exemplo, pode citar o dele, que é o mais famoso, que é o do Coronavirus Corpus” – como ele construiu a lista de palavras dele.

Você vai lá que você encontra como que ele encontrou os textos que são sobre o coronavírus.

A lista de palavras não é construída por frequência pura nem por glossários como critério único.

O critério metodológico é:

- aumentar o número de **verdadeiros positivos**;
- diminuir o número de **falsos negativos**.

O tamanho da lista é irrelevante.

8.1 Seleção inicial de termos

Os termos iniciais são selecionados a partir do conhecimento do domínio, por exemplo:

- climate change;
- climate emergency;
- global warming;
- climate crisis.

A legitimidade dessa escolha decorre da responsabilidade do pesquisador pelo desenho da pesquisa.

8.2 Validação empírica no Google Scholar

Para cada termo candidato:

- realizar uma busca no Google Scholar;
- examinar os 20 primeiros resultados;
- identificar quantos são falsos positivos;
- manter apenas termos com, no mínimo, 95% de verdadeiros positivos.

Não é necessário baixar os artigos nem realizar leitura aprofundada: a avaliação é feita a partir do título e do excerto do resumo.

8.3 Expansão iterativa por keywords

- Identificar artigos altamente citados;
- Extrair suas keywords;
- Examinar os 20 primeiros artigos que os citam;
- Extrair suas keywords;
- Incorporar novos termos se aumentarem verdadeiros positivos e reduzirem falsos negativos.

8.4 Filtragem final

- Eliminar termos que gerem muitos falsos positivos;
- Consolidar a lista final.

8.5 Regra de busca

As buscas são realizadas em condição de OU:

- termo A OU termo B OU termo C.
- NÃO EM CONJUNÇÃO RÍGIDA

9. Balanço entre Estratos

Cada estrato possui exatamente 200 sampling units.

Estratos com poucos textos (por exemplo, IPCC) são ampliados por meio de segmentação.

Textos longos da ONU também são segmentados.

Não se descarta material disponível.

10. Divisão entre LLMs

Os textos são divididos igualmente entre os modelos.

Mantém-se paridade humano/máquina.

Não há duplicação por tipo de prompt.

11. Limites Técnicos

Os limites de extensão de saída das LLMs são tratados como uma restrição técnica variável.

Reduções moderadas entre o tamanho do segmento humano e o tamanho do texto gerado são aceitáveis.

Discrepâncias drásticas não são.

12. Síntese do Desenho Metodológico

- Estratos definidos por ator social + registro;
- Sampling units segmentadas;
- Balanço de 200 unidades por estrato;
- Paridade humano/máquina;
- Prompt único detalhado;
- Reconstrução da cena;
- Controle da informação, não do discurso;
- Lista lexical construída por verdadeiros positivos e falsos negativos;
- Busca em OU;
- Validação empírica no Google Scholar.