

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO

**PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA
APLICADA E ESTUDOS DA LINGUAGEM**

Título do projeto: Modelagem de perfis sociais em sistemas de IA: Uma Análise Multidimensional Lexical de conversações humanas e geradas artificialmente

0. Resumo

O presente projeto tem por objetivo examinar a capacidade de modelos de Inteligência Artificial (IA) em simular identidades humanas específicas, notadamente de gênero, raça e personalidade, no âmbito da chamada Inteligência Artificial Social. Parte-se do pressuposto de que uma representação realista da interação social exige da IA mais do que a reprodução de um sujeito genérico; exige a incorporação de configurações variadas, ancoradas em dimensões social e psicologicamente marcadas. Para tanto, serão elaborados prompts contendo instruções explícitas que articulam traços sociais e psicológicos, os quais serão utilizados para orientar a geração de turnos de fala pela IA no contexto de conversações autênticas extraídas do British National Corpus. A IA assumirá o papel de um dos interlocutores humanos, produzindo enunciados em conformidade com os traços atribuídos. O corpus incluirá 4.900 textos, abrangendo produções humanas e sintéticas. A análise do corpus será conduzida por meio da Análise Multidimensional Lexical (AMDL), uma abordagem da Linguística de Corpus, que permitirá identificar dimensões latentes a partir da coocorrência estatística de palavras-chave. Serão examinadas as diferenças entre os perfis simulados e entre estes e as conversações humanas, com vistas a avaliar a consistência, os limites representacionais e os potenciais vieses na simulação humana operada pela IA.

1. Introdução

A investigação da simulação de identidades humanas por meio da Inteligência Artificial (IA), no âmbito da subárea conhecida como Inteligência Artificial Social (Dignum, 2018), fundamenta-se na premissa de que os sistemas de IA necessitam reproduzir uma variedade de características humanas fidedignamente. Nessa perspectiva, seria insuficiente emular um ser humano genérico na produção textual por meio de IA, já que os seres humanos variam entre si.

Nesse contexto, o projeto adotará uma perspectiva de ‘perfis sociais,’ isto é, de conjuntos de categorias específicas de traços humanos que a IA deve simular, dentre os quais se destacam as relacionadas a gênero (atributos estereotípicamente masculinos, femininos ou de pessoas trans), além de características psicológicas (tais como cordialidade e dominância) e de raça (negro, branco, etc.). Em nossa pesquisa, esses traços serão operacionalizados em prompts (instruções passadas aos sistemas de IA) que conterão referência explícita a determinadas características (gênero x, personalidade y, raça z). Além desses, empregaremos prompts genéricos, sem especificação de gênero, personalidade ou raça, a fim de capturarmos o estado padrão do raciocínio da IA. Essas configurações passadas à IA são chamadas aqui de ‘perfis.’ A partir da análise das respostas textuais da IA, poderemos comparar como a máquina se comporta sob cada uma desses perfis e definir qual configuração de gênero, personalidade e raça mais se assemelha a seu estado ‘natural,’ isto é, ‘default,’ o qual representa o seu conhecimento embutido na fase de treinamento, além de descobrir como a máquina representa cada uma das configurações humanas trabalhadas (combinações de raça, personalidade e gênero).

Assim, os prompts especificarão as condições de gênero, traços psicológicos e raça que a IA deverá adotar nos textos (por exemplo, ‘fale como uma mulher negra calma,’ ‘fale

como um homem branco agressivo'). As respostas textuais da IA serão baseadas em conversações humanas reais, previamente ocorridas entre interlocutores humanos. Ou seja, cada texto do corpus fonte corresponderá a uma conversa humana autêntica. A IA será instruída a substituir um dos falantes nessa conversa e a produzir turnos de fala consistentes com a 'persona' ou personagem atribuído a ela, na forma de um gênero, raça e caracterização psicológica (e.g. homem trans branco dominante), de tal forma que suas colocações na conversa sejam coerentes com as colocações dos demais interlocutores.

Dessa forma, a IA deverá participar dessas conversas mantendo coerência local com os turnos anteriores e consistência com os perfis sociais que lhe foram atribuídos. A partir desses diálogos modificados, será criado um corpus abrangendo configurações de gênero, raça e traços psicológicos.

As respostas produzidas pela IA serão submetidas à Análise Multidimensional Lexical, ou AMDL (Berber Sardinha & Fitzsimmons-Doolan, 2025), a qual possibilitará a identificação dos principais discursos construídos pela IA em cada condição. A AMDL é uma abordagem baseada em corpus que busca modelar os discursos presentes em um corpus por meio de análise estatística das correlações entre itens lexicais. Essas correlações dão vazão, por sua vez, a fatores (conjuntos de itens lexicais correlacionados entre os textos), os quais, uma vez interpretados qualitativamente, correspondem aos discursos colocados em jogo pelos interlocutores. Como se trata de uma abordagem que visa a detectar a variação entre textos, será possível verificar até que ponto há variação entre os textos produzidos pela IA nas diversas condições experimentais (combinações de gênero e personalidade).

Espera-se que os resultados obtidos forneçam uma compreensão empírica de como a IA social produz traços humanos específicos em contextos de interação. Além disso, pretende-se avaliar tanto a fidelidade quanto a flexibilidade das simulações da IA frente a características humanas, fornecendo assim uma base para entender as características humanas embutidas nos Grandes Modelos de Linguagem (LLMs, na sigla em inglês, Large Language Models). Tais modelos são vistos como um sistema fechado, opaco, cujo conhecimento embutido é impossível de acesso direto, sendo necessária a observação indireta de seu estado por meio de respostas a prompts específicos, como os que serão empregados nesta pesquisa.

A justificativa parte da necessidade de compreender como a IA representa o ser humano sob diferentes configurações sociais e psicológicas. Apesar dos avanços na IA gerativa, ainda é incerta a forma como esses sistemas reagem a categorias marcadas, como gênero, raça e personalidade. Modelos de linguagem treinados com grandes volumes de dados refletem vieses sociais e podem reproduzir estereótipos quando instruídos a simular pessoas específicas. Isso revela tanto limitações técnicas quanto ideológicas desses sistemas. Torna-se, portanto, necessário investigar empiricamente como a IA articula discursos associados a identidades socialmente sensíveis.

Desse modo, o objetivo geral da pesquisa é investigar como LLMs simulam traços humanos específicos, notadamente de gênero, raça e personalidade, em contextos de interação social simulada, a partir da geração de textos conversacionais. Os objetivos específicos são: (1) Determinar as dimensões lexicais que correspondem à interação conversacional gerada pela IA; (2) Identificar as características linguísticas predominantes associados a cada configuração; (3) Detectar vieses linguísticos na representação textual de perfis marginalizados de raça e gênero realizadas pela IA.

Com base nesses objetivos, as perguntas de pesquisa que guiarão a investigação são:

1. Quais dimensões emergem dos textos gerados pela IA?
2. Quais as semelhanças e diferenças linguísticas em relação às dimensões entre os perfis de gênero, raça e personalidade, bem como de seus cruzamentos?
3. Em relação às dimensões, há evidência de viés da IA na construção dos diálogos de grupos socialmente marginalizados (i.e. trans e negro)?

2. Fundamentação Teórica

2.1 Linguística de Corpus

A Linguística de Corpus (LC) é uma abordagem metodológica que se caracteriza pelo estudo empírico da língua a partir de grandes coleções estruturadas de textos autênticos, denominadas corpora (Berber Sardinha, 2004). Um corpus é tradicionalmente definido como uma compilação sistemática de textos naturais, que podem incluir dados escritos e falados, armazenados e analisados por meios computacionais. Embora existam diversas definições na literatura especializada, é possível identificar características comuns que definem o que se entende por corpus. Entre essas características destacam-se: o tamanho expressivo dos dados, ultrapassando geralmente o limite do processamento manual; a naturalidade dos textos selecionados, garantindo que sejam representativos do uso real da língua; a organização criteriosa e planejada das fontes textuais; e sua disponibilização em formato eletrônico, permitindo análises automáticas e interativas.

O uso de corpora para o estudo linguístico precedeu a própria utilização do termo ‘corpus’. Originalmente, pesquisadores como John Sinclair já se referiam às suas coleções textuais simplesmente como ‘textos’ ou ‘coleções de textos’ (Sinclair, 1966). Foi somente na década de 1960, com a introdução formal do Brown Corpus por Nelson Francis, que o termo ‘corpus’ passou a ser adotado regularmente em estudos linguísticos para designar tais conjuntos textuais sistematizados.

A LC não se configura como uma disciplina linguística autônoma, mas como uma abordagem ou perspectiva investigativa que atravessa diversas áreas da linguística aplicada e teórica (Berber Sardinha, 2004). A LC se apoia amplamente em ferramentas computacionais, permitindo análises sofisticadas e rápidas de grandes volumes de dados linguísticos. Por meio do uso combinado de técnicas quantitativas e qualitativas, a LC permite examinar padrões de uso linguístico, estabelecer generalizações fundamentadas em evidências empíricas e investigar aspectos da língua que seriam dificilmente observáveis sem auxílio tecnológico.

Além disso, a LC possibilita a análise de fenômenos linguísticos em diversos níveis, como o léxico, a gramática, o discurso e a pragmática, bem como contribui para o estudo de contextos específicos, como a comunicação acadêmica, jornalística, política, digital e de aprendizagem de línguas. Também exerce papel em áreas interdisciplinares, como humanidades digitais, linguística forense e comunicação em saúde. Dessa forma, a LC é uma abordagem que oferece recursos teóricos, metodológicos e tecnológicos para a investigação empírica da linguagem em contextos variados.

2.2 Características da produção textual da Inteligência Artificial

O raciocínio sintético da IA apresenta características que o distinguem do raciocínio humano, como a incongruência comunicativa: seus textos exibem desvios lexicogramaticais que afetam a naturalidade e função social da linguagem (Berber Sardinha, 2024a). Os textos artificiais são moldados por múltiplas dimensões funcionais e ideológicas, mas tendem ao hiperrealismo: acentuam traços comunicativos de modo exagerado, reduzindo a variabilidade típica da linguagem humana. São simulacros (Baudrillard, 1983), destituídos de experiência vivida, que adquirem aparência de autenticidade.

Além disso, os modelos não apenas reproduzem, mas também criam novas discursividades (Bender, Gebru, McMillan-Major, & Shmitchell, 2021), frequentemente distantes das práticas culturais de certos grupos. Representam um tipo humano idealizado — branco, letrado, do Norte Global — reforçando ideologias dominantes. A ausência de distinção entre registros textuais e a baixa documentação dos dados de treinamento comprometem a autenticidade contextual da linguagem gerada.

2.3 Análise Multidimensional

A Análise Multidimensional (AMD), concebida originalmente por Douglas Biber nos anos 1980 (Biber, 1988), constitui uma abordagem destinada à descrição da variação linguística em corpora. Desenvolvida em resposta à constatação de que as práticas linguísticas são moldadas por variáveis situacionais e funcionais, as quais geram variação sistemática entre os textos, essa abordagem visa a apreender padrões latentes de coocorrência de traços linguísticos que respondem por essa variação. No âmbito dessa concepção, o construto de registro designa variedades linguísticas associadas a contextos de uso e propósitos comunicativos distintos, tais como a ficção literária, a correspondência pessoal, escrita acadêmica e jornalística.

A fundamentação teórica da AMD repousa sobre três pilares metodológicos: (i) a unidade de análise é o texto completo, respeitando-se a integridade de sua composição interna; (ii) adota-se uma perspectiva abrangente e não reducionista da variação, incorporando um leque amplo de traços linguísticos que inclui marcas gramaticais, léxicas e discursivas; e (iii) pressupõe-se a existência de dimensões latentes que organizam o uso linguístico em contínuos funcionais e ideológicos, refletindo as exigências comunicativas dos contextos de produção textual. A abordagem privilegia a análise empírica de dados autênticos e visa a identificar padrões recorrentes com base em procedimentos estatísticos multivariados.

O principal procedimento estatístico é a Análise Fatorial, a qual permite extrair grupos de características linguísticas correlacionados, os quais, após interpretação qualitativa, revelam dimensões funcionais subjacentes ao uso linguístico. O procedimento resulta na redução de dezenas de variáveis observadas a um conjunto limitado de fatores interpretáveis, cada um representando um eixo funcional de variação. A interpretação dessas dimensões é orientada por princípios funcionais, de modo que cada dimensão é compreendida em termos do papel pragmático-discursivo desempenhado pelos traços que a compõem.

O estudo inaugural de Biber (1988), baseado em um corpus representativo da língua inglesa contemporânea, identificou cinco dimensões fundamentais: (1) envolvimento versus informação; (2) orientação narrativa; (3) referência explícita versus dependência do contexto situacional; (4) argumentatividade; e (5) densidade abstrata. Tais dimensões demonstraram poder discriminativo entre registros, evidenciando diferenças estatísticas entre, por exemplo, conversação espontânea e escrita acadêmica. Esse modelo se consolidou como uma referência metodológica para investigações interlingüísticas e interdisciplinares sobre variação textual.

A partir desse arcabouço inicial, a Análise Multidimensional evoluiu e diversificou-se em vertentes específicas. A AMDL constitui uma dessas vertentes (Berber Sardinha & Fitzsimmons-Doolan, 2025), priorizando exclusivamente unidades lexicais e suas distribuições semânticas e discursivas para a identificação de padrões ideológicos e posicionamentos em discursos especializados. Outras ramificações incluem a Análise Multidimensional Visual (Berber Sardinha, 2024b), voltada à investigação da comunicação visual por meio de ferramentas de visão computacional, e a Análise Multidimensional Multimodal (Delfino, Berber Sardinha, & Collentine, 2021), que abrange dados de vários modos semióticos, a fim de compreender a articulação entre múltiplos modos em práticas comunicativas.

Cada uma dessas extensões mantém a coerência epistemológica com a proposta original, ao mesmo tempo que amplia seu escopo empírico e analítico. A AMD fornece uma infraestrutura teórico-metodológica para o estudo das configurações linguísticas e discursivas que emergem em distintos contextos sociais. Sua aplicação permite descrever a variação registrada em corpora, por meio da aferição das pressões funcionais e ideológicas que motivam essa variação.

2.4 Linguística de Corpus e Inteligência Artificial

A relação entre LC e IA fundamenta-se nas semelhanças e complementaridades entre as duas áreas (Collentine & Berber Sardinha, *in press*). Historicamente, a IA evoluiu da tentativa de criar máquinas que imitassem processos cognitivos humanos, especialmente relacionados ao processamento da linguagem natural. Desde as primeiras abordagens simbólicas até os modernos modelos baseados em aprendizado profundo, a IA tem se beneficiado do uso de grandes conjuntos de dados linguísticos organizados em corpora, que são o foco da LC.

Atualmente, um ponto central na interseção entre IA e LC são os LLMs, que são treinados em conjuntos de textos a fim de serem capazes de gerar produções textuais próximas

ao padrão humano. No entanto, apesar de sua fluência, estudos baseados em corpus demonstram que os textos gerados por IA frequentemente apresentam diferenças linguísticas profundas em relação aos textos produzidos por humanos, particularmente no que se refere à coerência discursiva, uso de traços linguísticos específicos e adaptação a diferentes contextos comunicativos (Berber Sardinha, 2024a).

O presente projeto baseia-se em estudos em andamento na interface da AMDL com a IA, para identificar e comparar padrões linguísticos de textos sintéticos. Esses estudos indicam que os textos gerados por IA frequentemente apresentam discrepâncias em dimensões importantes, como grau de envolvimento, uso de recursos narrativos, referências explícitas e dependentes de contexto, além de capacidade de persuasão e nível de abstração (Berber Sardinha, 2024a).

2.5 IA Social

A fim de modelar aspectos sociais da IA, basear-nos-emos em estudos empíricos no âmbito da Psicologia Social que relatam como características humanas são representadas em sociedade. Relatamos, a seguir, estudos sobre estereotipia de gênero e raça conduzidos principalmente nos EUA, tendo em vista que os LLMs são produzidos com dados derivados essencialmente de produções textuais norte-americanas, sendo, portanto, mais provável que tenham adquirido estereótipos correntes naquele país.

Em relação ao gênero, Norton and Herek (2013) enfocam as atitudes de indivíduos heterossexuais em relação a pessoas transgênero. Metodologicamente, o estudo utilizou escalas de avaliação afetiva do tipo ‘termômetro’ (*feeling thermometer*), combinadas a medidas sociodemográficas, psicossociais e ideológicas. A partir de análise estatística de regressão, o estudo mostra que variáveis como autoritarismo, antigualitarismo e religiosidade mantiveram associação significativa com atitudes negativas em relação a pessoas transgênero.

O estudo, que utilizou dados de mais de dois mil respondentes, revelou que homens heterossexuais demonstram atitudes mais negativas em relação a pessoas transgênero do que suas contrapartes femininas. Entre os preditores mais robustos dessas atitudes desfavoráveis encontram-se a adesão a concepções binárias de gênero, níveis elevados de autoritarismo psicológico, alinhamento ideológico conservador e posicionamentos antigualitários.

Nesse sentido, a maneira como a IA representa identidades transgênero constitui um ponto de inflexão crítico para aferir o grau de fidelidade desses sistemas na replicação da diversidade humana, assim como seu potencial para reforçar ou subverter lógicas normativas. Tais achados são relevantes na medida em que sugerem que os LLMs podem ter incorporado representações linguísticas e sociais negativas a pessoas trans durante o treinamento. Em nossa pesquisa, ao induzirmos a IA a performarem personagens socialmente marcadas, como uma mulher trans, por exemplo, torna-se possível observar como essas tecnologias reproduzem, deslocam ou silenciam atributos identitários.

Ainda em relação ao gênero, Costa, Terracciano and McCrae (2001) enfocam traços de personalidade atribuídos a homens e mulheres, por meio de uma amostra de 23 mil

respondentes em 26 países. No domínio do neuroticismo, as mulheres consistentemente apresentaram escores mais elevados em facetas como ansiedade, depressão, autoconsciência e vulnerabilidade. Em relação à extroversão, as mulheres superaram os homens nas facetas de aconchego, gregarismo e emoções positivas, ao passo que os homens demonstraram escores superiores em assertividade e busca por emoções. Esses resultados sustentam a hipótese de uma estrutura dimensional interpessoal em que os homens se alinham mais a traços de dominação e excitação, enquanto as mulheres se orientam para o afeto interpessoal e responsividade emocional. O estudo relata variação entre respondentes de países diferentes (como EUA, Grécia, Bélgica, etc.). Porém, no geral, é possível detectar um padrão transcultural relativamente estável: mulheres tendem a ser associadas a características como neuroticismo, amabilidade e facetas afetivas da extroversão e abertura à experiência; homens, por sua vez, são mais associados a assertividade, busca por emoções e abertura a ideias. Esses padrões configuram-se em estereótipos que circulam na sociedade, sendo, portanto, potenciais traços formadores do raciocínio da IA que podem aflorar na sua interação com o ser humano.

Em relação à raça, há vários estudos empíricos que demonstram associação de estereótipos a grupos raciais. Por exemplo, Brigham (1971) consultou 200 respondentes brancos de nível universitário nos EUA e descobriu que suas atitudes perante a brancos e negros evidenciam um padrão claro: os brancos foram percebidos como mais competentes e racionais, enquanto os negros foram associados a traços ligados à emotividade, indisciplina e irresponsabilidade. Os negros foram mais frequentemente associados a traços como ‘musical,’ ‘atlético,’ ‘amistoso,’ ‘impulsivo,’ ‘gastador’ e ‘irresponsável.’ Por outro lado, traços positivos relacionados à competência e à disciplina, como ‘inteligente’ e ‘trabalhador,’ foram raramente atribuídos aos negros. Os brancos, ao contrário, foram mais frequentemente associados a traços positivos. A diferença percentual entre a frequência de atribuição dos traços para brancos e negros reforça essa tendência: ‘inteligente’ e ‘trabalhador’ apresentaram diferenças de 32 pontos percentuais a favor dos brancos. Traços negativos como ‘hostil,’ ‘impulsivo’ e ‘gastador,’ por sua vez, foram atribuídos com mais frequência aos negros.

Embora se observe certa mudança ao longo do tempo, os estereótipos raciais permanecem assimétricos. Katz and Braly (1933) evidenciaram que os negros eram amplamente associados a traços negativos, como ‘preguiçoso’ (75%) e ‘supersticioso’ (84%), enquanto os brancos recebiam traços positivos como ‘inteligente’ (48%) e ‘trabalhador’ (49%). Décadas depois, Maykovich (1971) constatou leve aumento na atribuição de traços positivos aos negros, como ‘inteligente’ (20%), mas o contraste entre os grupos seguiu presente. Nos anos 1980, Clark and Pearson (1982) indicaram certa atenuação dessas polarizações: ‘inteligente’ foi atribuído a 46% dos negros, e traços como ‘religioso’ passaram a ser mais igualmente distribuídos. Ainda assim, características estigmatizantes como ‘agressivo’ e ‘impulsivo’ continuaram mais associadas aos negros, enquanto os brancos mantiveram a predominância em traços de competência.

A comparação entre os três estudos revela que, apesar de pequenas mudanças ao longo do tempo, os estereótipos raciais nos Estados Unidos mantêm uma estrutura assimétrica. Os negros continuam sendo associados a traços ligados à fisicalidade, emotividade e impulsividade, enquanto os brancos são sistematicamente descritos como mais competentes, disciplinados e inteligentes. Essa estrutura discursiva permanece funcional à reprodução das hierarquias raciais no imaginário social americano.

3. Metodologia

Esta seção de metodologia apresenta os principais elementos da abordagem metodológica a ser empregada na pesquisa, especificamente o corpus fonte, o desenho de instrução para a IA (prompt design), a geração textual sintética, a construção do corpus e sua análise. O registro a ser investigado é a conversação informal, entendida como um tipo particular de interação verbal caracterizado por trocas espontâneas entre dois ou mais participantes, em contextos informais e não institucionalizados. Trata-se de interações em que os interlocutores se revezam livremente na fala, sem papéis predeterminados ou normas estruturadas por instituições. Essa forma de fala representa um polo mais cotidiano e descompromissado dentro do espectro mais amplo das práticas interacionais, conforme proposto por Hakulinen (2009).

Em relação ao corpus fonte, que representará a conversação humana, utilizaremos o British National Corpus (Love, Brezina, McEnery, Hawtin, Hardie et al., 2019), que comprehende conversações em inglês ocorridas na Grã-Bretanha nos anos 2010. Selecionaremos 100 conversações aleatoriamente do BNC e em seguida recolheremos os 30 primeiros turnos da conversação. Essa seleção é necessária tendo em vista a restrição de tamanho de texto de entrada (input) que os LLMs adotam, o que impede que conversações inteiras sejam inseridas no sistema.

Em relação à instrução (*prompting*), empregaremos um modelo de direcionamento baseado em oferecer ao LLM um roteiro que consiste em resumos de fala que o modelo deve transformar em diálogo (vide Quadro 1). Cada direção de fala resume brevemente o turno correspondente, a fim de orientar a IA na geração do diálogo humano.

| Conversação humana fonte | Direções de fala para IA |
|--|--|
| <u n='1' who='S0024'>an hour later <pause dur='short'/'> hope she stays down <pause dur='short'/'> rather late</u> | <u n='1'>Discussing being late.</u> |
| <u n='2' who='S0144'>well she had those two hours earlier</u> | <u n='2'>Mentioned previous plans.</u> |
| <u n='3' who='S0024'>yeah I know but that's why we're an hour late isn't it? <pause dur='long'/'> mm <pause dur='short'/'> I'm tired now</u> | <u n='3'>Explaining the delay.</u> |
| <u n='4' who='S0144'><vocal desc='laugh'/'></u> | <u n='4'>Laughed.</u> |
| <u n='1' who='S0024'>an hour later <pause dur='short'/'> hope she stays down <pause dur='short'/'> rather late</u> | <u n='5'>Asking about texting.</u> |

Quadro 1: Amostra de direções de fala para geração de diálogos por IA baseadas em conversação humana.

A geração sintética será realizada por dois modelos, o GPT o3 e Grok 3.0, dois LLMs de alta capacidade de raciocínio. A IA será instruída a desempenhar duas características raciais (negro, branco), três gêneros (feminino, masculino, trans) e quatro características psicológicas (a serem determinadas no andamento do projeto). A geração dos textos se dará por meio de API, utilizando scripts em Python a serem desenvolvidos pelo grupo de pesquisa. Desse modo, o corpus assumirá a feição do desenho mostrado no Quadro 2.

| | Fonte humana | LLMs | Perfis raciais | Perfis de gênero | Perfis psicológicos | Total |
|--------|--------------|------|----------------|------------------|---------------------|-------|
| Textos | 100 | 2 | 2 | 3 | 4 | 5000 |

Quadro 2: Desenho do corpus.

Por fim, a análise dos 5000 textos do corpus (4800 textos gerados pela IA relacionados aos perfis, 100 gerados por IA sem um perfil definido [perfil genérico] e 100 produzidos pelos seres humanos) se dará por meio da AMDL, que consiste nos passos descritos a seguir. Os textos serão anotados morfossintaticamente e lematizados por meio da ferramenta TreeTagger. Para cada lema, será registrado o número de textos em que ocorrer. Tais contagens serão utilizadas para análise de palavras-chave com o objetivo de identificar os lemas mais distintivos de cada perfil (sintético e humano). Um lema será considerado palavra-chave caso atenda simultaneamente a três critérios: (i) ocorrer em pelo menos 2% dos textos do corpus; (ii) apresentar distribuição proporcionalmente maior no subcorpus-alvo (composto pelos textos de um determinado perfil) do que no subcorpus de comparação (textos de todos os demais perfis); e (iii) atingir valor de log-verossimilhança igual ou superior a 3,84, equivalente ao nível de significância $p < 0,05$. Somente os lemas que satisfizerem os três critérios serão classificados como palavras-chave e retidos para as etapas subsequentes da análise.

Em seguida, as contagens de lemas serão alvo de análise fatorial, que detectará os agrupamentos correlacionados desses itens, os quais serão interpretados, passando a ser dimensões de variação. Serão conduzidos testes estatísticos com base nos escores dos textos em cada dimensão a fim de verificar as diferenças e semelhanças entre os perfis de raça, gênero e de personalidade da IA (bem como seus cruzamentos) entre si e em relação aos dos seres humanos.

4. Cronograma

| Semestre | Disciplinas | Leituras | Coleta do corpus | Análise do corpus | Redação da dissertação | Qualificação | Depósito |
|----------|-------------|----------|------------------|-------------------|------------------------|--------------|----------|
| 2/2025 | X | X | X | | | | |
| 1/2026 | X | X | X | X | X | | |
| 2/2026 | | | | X | X | X | X |

5. Referências bibliográficas

- Baudrillard, J. (1983). *Simulations*. Los Angeles: Semiotext(e).
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* Paper presented at FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,
- Berber Sardinha, T. (2004). *Linguística de Corpus*. São Paulo: Manole.
- Berber Sardinha, T. (2024a). AI-generated vs human-authored texts: A Multidimensional comparison. *Applied Corpus Linguistics*, 4(1), 100083. <https://doi.org/https://doi.org/10.1016/j.acorp.2023.100083>
- Berber Sardinha, T. (2024b). Exploring multimodal corpora in the classroom from a multidimensional perspective In P. Crosthwaite (Ed.), *Corpora for Language Learning: Bridging the Research-Practice Divide* (pp. 25-36). Abingdon: Routledge.
- Berber Sardinha, T., & Fitzsimmons-Doolan, S. (2025). *Lexical Multidimensional Analysis: Identifying Discourses and Ideologies*. Cambridge: Cambridge University Press.
- Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.
- Brigham, J. C. (1971). Ethnic stereotypes. *Psychological Bulletin*, 76(1), 15–38.
- Clark, M. L., & Pearson, W., Jr. (1982). Racial stereotypes revisited. *International Journal of Intercultural Relations*, 6, 381–393.
- Collentine, J., & Berber Sardinha, T. (in press). Artificial Intelligence and Corpus Linguistics. In C. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (2nd ed.). Malden, MA: Wiley.
- Costa, P. T. J., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, 81(2), 322–331. <https://doi.org/10.1037/0022-3514.81.2.322>
- Delfino, M. C. N., Berber Sardinha, T., & Collentine, J. G. (2021). *Tipologia multidimensional multimodal big data da música pop em inglês*. LAEL, PUCSP.
- Dignum, V. (2018). *Responsible Artificial Intelligence: Designing AI for Human Values* (1st ed.). Cham: Springer. <https://doi.org/10.1007/978-3-030-30371-6>
- Hakulinen, A. (2009). Conversation types. In S. D'Hondt, J.-O. Östman, & J. Verschueren (Eds.), *The Pragmatics of Interaction* (pp. 55-65). Amsterdam: John Benjamins.
- Katz, D., & Braly, K. (1933). Racial stereotypes of 100 college students. *Journal of Abnormal and Social Psychology*, 28, 280–290.
- Love, R., Brezina, V., McEnery, T., Hawtin, A., Hardie, A., & Dembry, C. (2019). Functional variation in the Spoken BNC2014 and the potential for register analysis. *Register Studies*, 1(2), 296-317. <https://doi.org/https://doi.org/10.1075/rs.18013.lov>
- Maykovich, M. (1971). Changes in racial stereotypes among college students. *Human Relations*, 24(5), 371–386.
- Norton, A. T., & Herek, G. M. (2013). Heterosexuals' attitudes toward transgender people: Findings from a national probability sample of U.S. adults. *Sex Roles*, 68, 738–753. <https://doi.org/10.1007/s11199-011-0110-6>
- Sinclair, J. M. (1966). Beginning the study of lexis. In C. E. Bazell (Ed.), *In Memory of J R Firth* (pp. 410-430). London: Longman.

