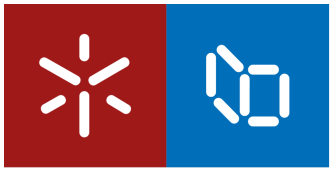


Universidade do Minho
Escola de Letras, Artes e Ciências Humanas

Relatório Análise e Visualização de Dados
Baú de Família

Trabalho realizado por:

- Ana Margarida Cardoso, pg57937;
- Beatriz Oliveira, pg56347.

Índice

Introdução

O presente relatório tem como objetivo apresentar e analisar o conteúdo de quatro entrevistas realizadas no âmbito de um projeto sobre memórias de infância. As entrevistas foram conduzidas com dois grupos distintos de participantes:

- **Carina Faria** e **João Silva**, entrevistados por **Beatriz Oliveira**;
- **Catarina Miranda** e **Catarina Pereira**, entrevistadas por **Ana Margarida Cardoso**.

Todos os participantes aceitaram colaborar gentilmente neste trabalho académico, partilhando as suas memórias mais significativas da infância.

A análise dos dados baseou-se em transcrições feitas em formato .md (Markdown), complementadas por diversas técnicas de Processamento de Linguagem Natural (PLN). Entre as abordagens utilizadas destacam-se: a extração de entidades nomeadas, análise de palavras-chave e *chunks*, frequência de termos e avaliação do sentimento associado ao discurso de cada participante.

Este trabalho visa, assim, comparar não só os dados das entrevistas individuais, mas também refletir sobre os aspetos emocionais e linguísticos que emergem dos relatos pessoais dos quatro entrevistados.

Contexto das Entrevistas

As entrevistas foram realizadas em dois momentos e contextos distintos:

- **Carina Faria** e **João Silva** são amigos da entrevistadora **Beatriz Oliveira** desde o ensino secundário. Mostraram-se disponíveis e entusiasmados por partilhar as suas vivências. As entrevistas decorreram em ambiente informal, foram gravadas em áudio e posteriormente transcritas. O conteúdo final foi revisto manualmente para assegurar fidelidade à linguagem original.

- **Catarina Miranda** e **Catarina Pereira** foram entrevistadas por **Ana Margarida Cardoso**, também participante neste projeto. Ambas têm uma ligação pessoal à entrevistadora: uma foi sua colega no ensino secundário e a outra foi sua explicadora. Essa relação de proximidade permitiu criar um ambiente de confiança, proporcionando relatos autênticos, espontâneos e afetivos sobre as suas experiências de infância.

A junção destes dois conjuntos de entrevistas enriquece a diversidade do corpus analisado, permitindo uma leitura comparativa mais profunda sobre os diferentes estilos de discurso, níveis de expressividade emocional e conteúdos temáticos abordados.

Quantificação dos Dados

As entrevistas transcritas variam significativamente em extensão. As entrevistas A1 e A2, obtidas automaticamente, contêm, respetivamente, 115 linhas e 1691 palavras (A1), e 72 linhas e 850 palavras (A2). Já as outras duas entrevistas apresentam maior volume, a entrevista B1 contém 176 linhas e 3241 palavras, enquanto que a entrevista B2 apresenta 156 linhas e 2676 palavras. Estes valores ilustram não só diferenças nos estilos e níveis de detalhe dos relatos, mas também refletem as distintas metodologias de transcrição e recolha de dados adotadas nos dois conjuntos de entrevistas.

Link Github: https://github.com/paratrabalhos20232024/MuseuPessoa2425

O nosso trabalho está na pasta Trabalho_Final

Análise e Comparação

1. Estilo e Estrutura Textual

A1 – Estilo Narrativo e Pessoal

O A1 destaca-se por ter uma estrutura frásica mais desenvolvida e próxima da oralidade real. As frases são completas, há uso de perguntas e respostas, interações diretas e linguagem informal. É um texto emocionalmente rico, que reproduz o tom de uma conversa espontânea, como se fosse transcrição de uma entrevista vivida.

A2 – Estilo Descritivo e Contido

O A2, embora baseado nas mesmas experiências, é uma versão mais "filtrada". As frases são mais curtas, mais focadas em nomes e adjetivos do que em verbos e ações. O discurso é contido, mais formal, com menor presença de marcas da oralidade. Parece haver uma intenção de "resumo" ou distanciamento emocional.

B1 – Carina: Espontâneo e Afetivo

Carina fala com naturalidade, com muitas expressões do dia a dia ("pronto", "então"), o que torna o discurso muito acessível e realista. Utiliza frequentemente o possessivo ("a minha mãe", "o meu irmão"), revelando um discurso centrado na relação com os outros. O vocabulário é concreto e sensorial.

B2 – João: Reflexivo e Organizado

João tem um estilo mais limpo, com menos marcas de oralidade. Mostra uma organização clara do pensamento, com descrições detalhadas, mas menos emocionais. As suas memórias são muitas vezes mediadas por objetos culturais (jogos, televisão), sugerindo uma ligação mais racional e narrativa aos eventos.

2. Vocabulário e Palavras-Chave

A1 e B1: O vocabulário foca-se em experiências diretas e relações familiares. Destacam-se palavras como "irmão", "pais", "casa", "avó", "mãe". São textos íntimos e sensoriais, com muitos termos associados ao toque, cheiro, rotinas e sentimentos.

A2 e B2: Incluem mais termos abstratos ou descritivos como "memórias afetivas", "influência", "sensação", "lembrança". Aparecem também referências a objetos externos como "Playstation", "cartas", "canal de TV", o que mostra uma memória mais mediada pela cultura e tecnologia.

3. Makefile

Para facilitar a execução automática das várias análises aplicadas às entrevistas, foi criada uma `Makefile` que organiza e encadeia todas as tarefas do projeto. A `Makefile` permite executar comandos como extração de entidades, análise de sentimentos, gerar gráficos e cálculos de frequência com um simples `make`. Esta abordagem garante uma maior organização, poupa tempo e assegura que todas as etapas são executadas de forma consistente em todos os ficheiros.

4. Markdown

As entrevistas foram transcritas e estruturadas em ficheiros `Markdown (.md)`, um formato leve que permite integrar texto, metadados, imagens e marcações de fala com clareza. Cada bloco de diálogo foi etiquetado com `<entrevistador>` e `<entrevistado>`, o que facilitou a identificação automática de quem fala em cada momento. Esta estrutura organizada foi essencial para aplicar técnicas de Processamento de Linguagem Natural (PLN) de forma precisa.

5. Conll

Inicialmente, geramos um ficheiro em formato **Conll** para cada entrevista, o que nos permitiu obter anotações linguísticas palavra a palavra, incluindo classes gramaticais e relações sintáticas. No entanto, como este formato inclui **todas as palavras do texto**, percebemos que a análise se tornava demasiado extensa e, por vezes, pouco relevante para os nossos objetivos. Por isso, optámos por **focar a nossa atenção nas entidades mais específicas e significativas**, como nomes próprios de pessoas e locais, que melhor revelam os temas e o contexto emocional de cada discurso.

6. Entidades

A análise das quatro entrevistas (A1, A2, B1 e B2) teve como objetivo identificar os protagonistas, locais e marcas culturais mencionados, observando simultaneamente a qualidade e coerência da anotação das entidades nomeadas (NER – *Named Entity Recognition*). As entrevistas foram analisadas com recurso a ferramentas automáticas de NER.

Estrutura e Conteúdo

As entrevistas **A1 e A2** apresentam uma extração automática de entidades, com diferentes categorias sintáticas atribuídas a nomes próprios, locais, marcas e termos ambíguos. Já as entrevistas **B1 (Carina)** e **B2 (João)** centram-se na dimensão narrativa das memórias, identificando com clareza os protagonistas, locais de referência e elementos simbólicos da infância.

Principais Resultados

- **Entidades de Pessoa**:
  A1 contém nomes como *Margarida*, *Catarina* e *Manténs*, com variações compostas e frequências anormalmente elevadas (superiores a 500.000), o que indica erros de contagem. A2 praticamente não apresenta pessoas. Por contraste, B1 identifica pessoas reais próximas (irmão, professora, entrevistadora), e B2 apenas menciona *João*.

- **Locais (LOC)**:
  A1 e A2 reconhecem lugares como *Braga*, *Jardim* e *Concelho de Barcelos*, mas também cometem erros, como considerar verbos conjugados (ex.: *Escolhíamos*) como locais. Em B1 e B2, os locais são contextualizados e geograficamente coerentes, como *Ronfe*, *Famalicão*, *Mogege* e *Póvoa de Varzim*.

- **Organizações e Marcas (ORG)**:
  A2 identifica corretamente *Nivea*, enquanto B2 refere marcas significativas da infância como *Playstation 2*, *WWE* e *Yu-Gi-Oh!*. Em A1, não se identificam organizações relevantes.

- **Entidades Mistas (MISC) e Outros**:
  A1 e A2 incluem termos como *Páscoa*, *Verão*, *Macaca* e *Olá*, com classificações ambíguas. *Olá* é erradamente identificado como local e interjeição. Em B1 e B2, surgem entidades como *Mãe Natureza*, *Natal* e *Histórias*, com valor simbólico e afetivo, integradas no discurso.

Qualidade e Coerência da Anotação

- As **entrevistas automáticas (A1 e A2)** apresentam problemas como:
    - Classificação incorreta de verbos como entidades nomeadas;
    - Frequências anormalmente altas que indicam falhas técnicas;
    - Mistura entre nomes reais, termos simbólicos e erros de leitura;
    - Funções sintáticas atribuídas corretamente em alguns casos (ex.: *Nivea/appos*), mas com muitos ruídos.
- As **entrevistas manuais (B1 e B2)** revelam:
    - Elevada coerência entre entidades e discurso;
    - Narrativas pessoais (B1) ou culturais (B2) bem contextualizadas;
    - Clareza na identificação dos protagonistas e dos espaços relevantes.

Aspeto	A1	A2	B1 (Carina)	B2 (João)
Foco do discurso	Memórias ligadas a locais e elementos simbólicos; linguagem mais descritiva	Marcas e termos culturais; uso menos pessoal, mais objetivo	Relato pessoal e emocional, centrado na família, escola e espaços locais	Narrativa focada em entretenimento e referências culturais (jogos, TV)
Entidades de Pessoa	Margarida, Catarina, Catarina Pereira, Manténs (alguns	Nenhuma pessoa claramente identificável	Carina, Diogo, Carlos de Castro,	João

	com frequências irreais)		Beatriz Oliveira	
Locais (LOC)	Jardim, Solei, Concelho de Barcelos, Braga (com erros como "Escolhíamos" marcado como LOC)	Braga, (menção de outros locais irrelevantes ou ruidosos)	Ronfe, Joane, Póvoa de Varzim, Hospital de Famalicão, Portugal	Mogege, Rego, Famalicão
Organizações/Marcas (ORG)	Nenhuma relevante	Nivea (bem anotada, coerente)	Nesquik	Playstation 2, WWE, Yu-Gi-Oh!
Entidades Mistas (MISC)	Páscoa, Verão, Macaca, Olá (ambíguas ou mal classificadas)	Páscoa, Troco (Troco como ROOT é erro), Nasci (classificado como nome)	Mãe Natureza, Natal	Natal, Histórias (canal)

Conclusões

- As entrevistas A1 e A2 **não coincidem diretamente** com B1 ou B2, mas há aproximações temáticas:
    - **A1 aproxima-se de B1**, pelo uso de locais e símbolos do quotidiano.
    - **A2 aproxima-se de B2**, pelas referências a marcas e elementos culturais.

7. Temas e Foco das Narrativas

Tema	**A1**	**A2**	**B1**	**B2**
Infância e brincadeiras	Muito presentes	Mais simbólicas	Centrais e detalhadas	Centrais e ligadas a objetos
Família e afetos	Fortemente presente	Reduzido, indireto	Muito marcante (pais, avó)	Positivo e estável
Escola	Relatos variados	Pouca presença	Ambivalente (negativa e neutra)	Crítica leve, mas controlada
Emoções fortes	Sim, com variações	Poucas, mais neutras	Sim, oscilação afetiva clara	Estáveis, com distanciamento
Cultura popular	Pouca referência	Ausente	Referência a brinquedos (Winx)	Forte (jogos, WWE, media)

8. Análise de Sentimentos

Iniciámos a análise de sentimentos com base nos parágrafos extraídos do nosso ficheiro Markdown, onde a entrevista estava estruturada em XML. Utilizámos apenas as respostas do entrevistado (conteúdo dentro da tag `<interviewee>`). No entanto, percebemos que esta abordagem não era suficientemente específica, pois alguns parágrafos eram demasiado curtos. Por isso, optámos por dividir o texto em conjuntos de palavras com dimensão mais uniforme. A partir dessa nova segmentação, criámos uma tabela que serviu de base para a construção dos gráficos de sentimento.

A1

- Polaridade média positiva, mas com variações maiores.
- Blocos com forte carga emocional — tanto positiva (nostalgia, carinho), como negativa (tristeza, perda).

A2

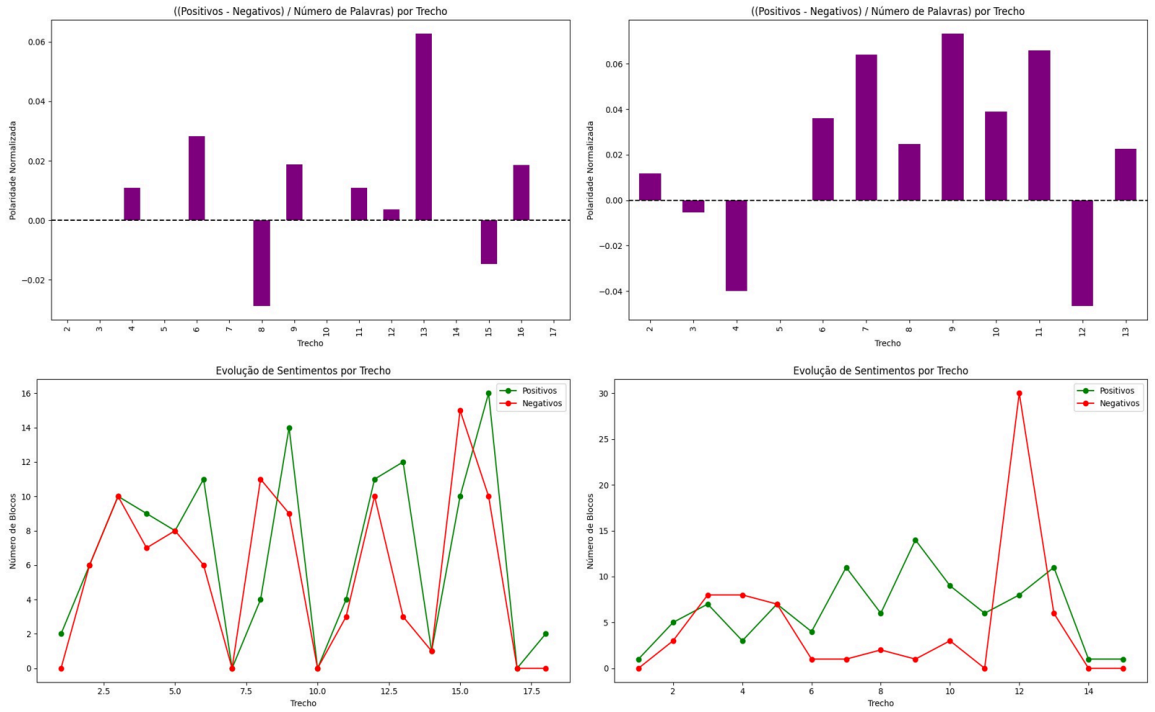
- Discurso mais "apagado" emocionalmente.
- Mais blocos com sentimento negativo ou neutro, e quase sem picos positivos.

B1

- Muito expressiva emocionalmente. Fala abertamente de alegria, tristeza, raiva e ternura.
- Há humor involuntário, espontaneidade e um tom muito humano.

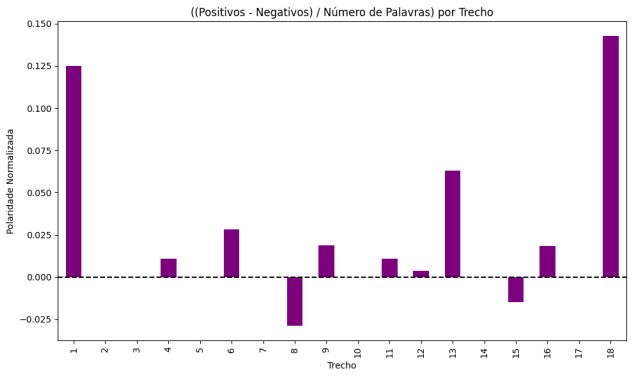
B2

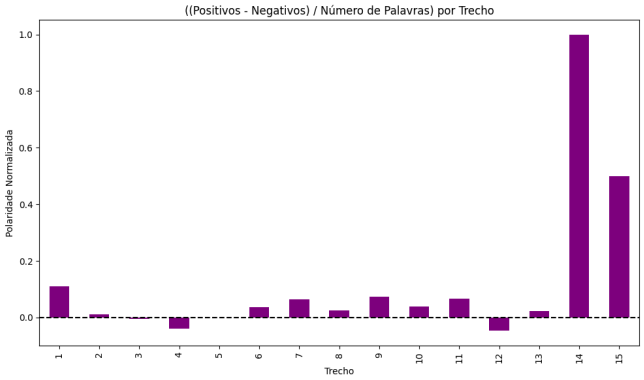
- Tom mais constante e positivo.
- Menos picos emocionais, mas narrativa igualmente afetiva. Sente-se nostalgia, com um certo recato e sensibilidade equilibrada.



No caso das entrevistas B1 e B2, foi necessário gerar outros gráficos de barras, por não ser possível interpretar corretamente os dados. Nos trechos em que havia agradecimentos ou despedidas, o número de palavras era muito baixo, o que distorcia a análise: esses segmentos acabavam por parecer ou muito positivos ou muito negativos, dificultando a visualização comparativa com os outros trechos, que eram mais longos.

Abaixo apresento os gráficos de barras originais (B1 e B2, respetivamente):





9. Modos e Tempos Verbais

Procurámos observar como cada entrevistado recorre ao indicativo, ou subjuntivo e ao condicional, bem como à variação temporal (passado, presente, futuro), de forma a compreender estilos narrativos e padrões discursivos. As diferenças no uso verbal ajudam a revelar o grau de espontaneidade, reflexão ou planeamento nas falas, bem como a forma como os entrevistados se posicionam em relação aos acontecimentos relatados.

Tempo/ Modo	**A1**	**A2**	**B1**	**B2**
Indicativo	Uso intensivo	Reduzido	Muito presente	Reduzido, mas variado
Pretérito imperfeito	Frequente (hábitos)	Menos marcado	Muito frequente	Presente, com variações
Subjuntivo / Condicional	Quase ausentes	Quase ausentes	Pouco usados	Começa a usar mais
Presente / Futuro	Simples e direto	Quase inexistente	Presente informal	Futuro/condicional leve

10. Coesão e Ligação entre Ideias

- **A1 e B1**: Ligação forte entre frases, muitos conectores como “porque”, “então”, “por isso”, o que dá fluidez.

- **A2 e B2**: Menor número de conectores, ideias menos encadeadas, estilo mais direto ou implícito. Isto pode indicar:
    - maior maturidade (menos dependência de conectores);
    - ou menor fluência discursiva (frases mais isoladas).

Comparação Geral

Dimensão	**A1**	**A2**	**B1**	**B2**
Estilo	Narrativo e espontâneo	Descritivo e controlado	Emocional e familiar	Nostálgico e racional
Complexidade	Média-alta	Média-baixa	Média	Média
Envolvimento emocional	Elevado	Reduzido	Muito elevado	Moderado e sereno
Oralidade	Forte	Fraca	Muito forte	Média
Coesão	Fluída	Fragmentada	Articulada	Contida
Vocabulário	Simples e direto	Abstrato e adjetival	Familiar, sensorial	Cultural e variado

11. Rácio de Ocorrência

O *rácio* entre a frequência observada de uma palavra e a sua frequência esperada segundo um corpus de referência permite identificar os termos mais distintivos e reveladores do discurso de cada entrevistado.

A1 – Entrevista centrada na autorreflexão e estrutura formal

Palavras com rácios mais elevados:

- *entrevistadora* (3773.9), *escolheria* (516.2), *Entrevistado*, *confiança*, *Solei*

Interpretação:
O discurso da A1 é altamente **metarreflexivo e consciente do contexto da entrevista**. Os termos com rácios elevados indicam uma entrevistada que fala sobre o seu papel, escolhas e valores. A linguagem é formal, estruturada e destaca a importância do momento da entrevista em si.

A2 – Discurso afetivo e nostálgico sobre a infância

Palavras com rácios mais elevados:

- *nadar* (1006.0), *divertir* (430.8), *deixavam*, *cresceste*, *Víamos*

Interpretação:
A2 apresenta uma narrativa **emocional e sensorial**, com foco na infância e nas memórias positivas. Os verbos conjugados no passado e a presença de atividades prazerosas refletem uma ligação afetiva com o passado. A linguagem é pessoal, íntima e com forte carga emocional.

B1 – Vocabulário familiar e emocional (Carina)

Palavras com rácios mais elevados:

- *Carina* (72.6), *infância* (11.85), *brincadeira* (14.62), *irmão*, *avó*, *mãe*, *pai*

Interpretação:
Carina apresenta um discurso **familiar, sensorial e marcado pela emoção**. As palavras com rácios elevados remetem a relações próximas e memórias concretas como *neve*, *sopa*, *café*,

berros. A expressão de sentimentos como *super chateada* e *feliz* reforça o tom vivencial e íntimo da entrevista.

B2 – Memória cultural e social (João)

Palavras com rácios mais elevados:

- *gostar* (17.97), *recordar* (14.99), *Playstation* (38.93), *Yu-Gi-Oh* (44.46), *turma*

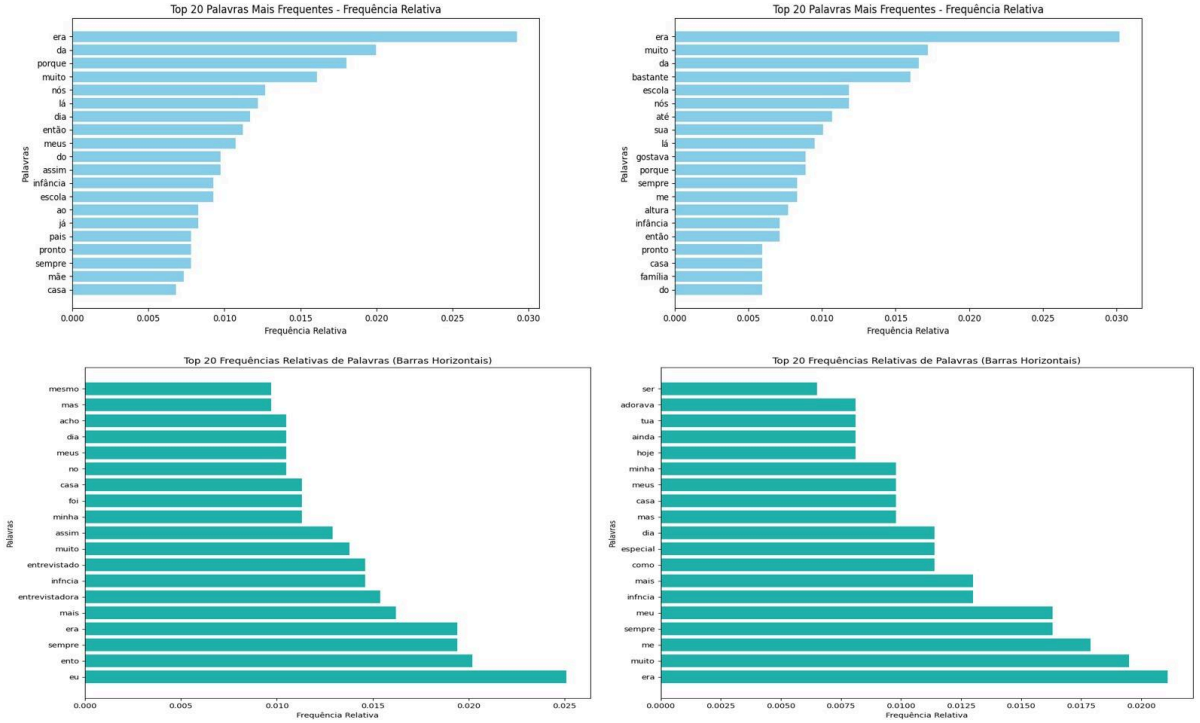
Interpretação:
João apresenta um discurso mais orientado para a **memória mediada por elementos culturais**, com destaque para brinquedos, jogos e referências sociais. Os termos evidenciam um foco na infância urbana, com ligações a colegas e primos. A linguagem é descritiva e marcada por **recordações conscientes e nostálgicas**.

Comparação geral

Entrevista	**Estilo discursivo**	**Palavra com maior rácios**	**Valor do rácios**	**Foco principal**
A1	Formal, reflexivo	*entrevistadora*	3773.90	Papel na entrevista, estrutura
A2	Emocional, narrativo	*nadar*	1006.04	Memórias da infância e sentimentos
B1	Regional, familiar	*Ronfe*	339.12	Vida quotidiana e relações locais
B2	Cultural, coloquial	*conheci*	258.16	Cultura jovem e relação social

12. Palavras Frequentes

Os gráficos a seguir apresentam as 20 palavras mais frequentes em cada entrevista, refletindo temas dominantes, estilo narrativo e nível de envolvimento emocional. A esse levantamento, somamos agora a análise das **interjeições e expressões marcadamente orais**, que ajudam a caracterizar o tom, a espontaneidade e o ritmo de cada fala.



B1 (superior esquerdo)

Além de um vocabulário fortemente centrado em memórias pessoais e familiares, B1 destaca-se pelo uso **muito expressivo de interjeições**:

- **“então” (24x)**, **“pronto” (16x)**, **“assim” (20x)**, **“bem” (9x)** e **“realmente” (8x)**.

- Outras expressões como **“tipo”**, **“ok”**, **“ah”**, **“olha”** e **“digo”** surgem pontualmente.
  Esse uso acentuado contribui para um discurso espontâneo, envolvente e muito marcado por marcas da oralidade. A presença dessas interjeições reforça a naturalidade e a emoção com que as memórias são partilhadas.

B2 (superior direito)

O discurso de B2 mantém o tom afetivo e reflexivo, mas com **menos interjeições do que B1**:

- Destaque para **“então” (12x)**, **“pronto” (10x)**, **“bem” (6x)** e **“realmente” (5x)**.

- Interjeições como **“tipo”**, **“claro”**, **“ok”** e **“logo”** aparecem com menor frequência.
  Esse uso mais moderado mantém certa espontaneidade, mas com maior contenção emocional e organização narrativa. As interjeições funcionam aqui mais como marcadores de discurso do que como traços emocionais intensos.

A2 (inferior direito)

A entrevista A2 apresenta o **uso mais reduzido de interjeições entre todas**:

- Apenas **“bem” (3x)** foi identificada.
  Esse padrão confirma a impressão de um discurso mais neutro, distanciado e pouco expressivo emocionalmente, refletido também na análise de sentimentos e na escolha vocabular.

A1 (inferior esquerdo)

Já A1 mostra um uso **acentuado e variado de interjeições**, próximo ao observado em B1:

- **“então” (25x)**, **“assim” (16x)**, **“coisa” (8x)**, **“bem” (5x)**, **“ah” (3x)** e **“tipo” (3x)**.

- A variedade e frequência dessas expressões revelam uma fala envolvida, espontânea e emocionalmente marcada, com forte presença de linguagem informal e oralizada.

Desta forma…

O uso de interjeições confirma e complementa os padrões observados na análise de frequência lexical e sentimentos. **B1 e A1** revelam discursos mais vivos, emocionais e espontâneos, enquanto **B2** equilibra afetividade com organização narrativa. **A2**, por outro lado, apresenta um estilo mais contido, reflexivo ou até mesmo retraído, com pouca marca de oralidade.

Conclusão

No final deste projeto, cada uma de nós compilou um dataset com todos os resultados obtidos, reunindo os dados tratados ao longo da análise. Este repositório constitui um contributo relevante, não só como síntese do trabalho desenvolvido, mas também como ponto de partida para futuras investigações na área.

A análise das entrevistas permitiu compreender a diversidade das memórias de infância e as diferentes formas como estas são expressas linguisticamente. Os quatro participantes — Carina, João, Catarina Miranda e Catarina Pereira — partilharam experiências únicas, revelando distintos estilos discursivos, graus de emocionalidade e focos temáticos. Enquanto alguns relatos evidenciaram uma forte carga afetiva e espontaneidade, outros demonstraram maior organização e reflexão, o que enriqueceu a comparação entre os discursos.

A utilização de técnicas de Processamento de Linguagem Natural (PLN) revelou-se essencial para identificar padrões lexicais, sintáticos e emocionais, permitindo uma análise mais sistemática e aprofundada.

Em suma, este trabalho evidenciou o papel fundamental da linguagem na reconstrução da memória pessoal. Recordar a infância revelou-se um exercício não só de rememoração, mas também de identidade, de expressão e de partilha individual — aspectos que, em conjunto, reforçam a importância da linguagem enquanto ferramenta de construção e comunicação da experiência humana.

Referências Bibliográficas

- Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). *Array programming with NumPy*. *Nature*, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2
  (para a biblioteca NumPy)

- Hunter, J. D. (2007). *Matplotlib: A 2D Graphics Environment*. *Computing in Science & Engineering*, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55
  (para a biblioteca Matplotlib)

- McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. In *Proceedings of the 9th Python in Science Conference*, 56–61. https://pandas.pydata.org/
  (para a biblioteca pandas)

- Montani, I., & Honnibal, M. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. To be published. Consultado em https://spacy.io/
  (para a biblioteca spaCy)

- Python Software Foundation. (s.d.). *Python 3.11.3 documentation*. Consultado a 1 de junho de 2025 em https://docs.python.org/3/
  (para Python e os módulos da biblioteca padrão: sys, os, re)

- Wolf, T., Debut, L., Sanh, V., et al. (2020). *Transformers: State-of-the-Art Natural Language Processing*. In *Proceedings of the 2020 Conference on EMNLP: System Demonstrations*, 38–45. https://huggingface.co/transformers
  (para a biblioteca Transformers da Hugging Face)