UNIVERSIDADE FEDERAL DE MINAS GERAIS
Faculdade de Letras
Programa de Pós-Graduação em Estudos Linguísticos

Ana Luiza de Souza Couto

"CEBOLA" OU "SEBOLA"?:
a evidência de múltiplos padrões no processo de escrita de grafemas concorrentes irregulares

Belo Horizonte – MG
2024

Ana Luiza de Souza Couto

"CEBOLA" OU "SEBOLA"?:
a evidência de múltiplos padrões no processo de escrita de grafemas concorrentes irregulares

Tese apresentada ao Programa de Pós-Graduação em Estudos Linguísticos da Faculdade de Letras da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Estudos Linguísticos.

Área de Concentração: Linguística Aplicada. Linha de Pesquisa: Ensino do Português. Orientadora: Prof.ª. Dra. Daniela Mara Lima Oliveira Guimarães.

Belo Horizonte – MG

2024

Couto, Ana Luiza de Souza.
C871c "CEBOLA" ou "SEBOLA"? [manuscrito] : a evidência de múltiplos padrões no processo de escrita de grafemas concorrentes irregulares / Ana Luiza de Souza Couto. – 2024.
1 recurso online (206 f. : il., tabs., grafs. (algumas color.)) : pdf.

Orientadora: Daniela Mara Lima Oliveira Guimarães.

Área de concentração: Linguística Aplicada.

Linha de Pesquisa: Ensino de Português.

Tese (doutorado) – Universidade Federal de Minas Gerais, Faculdade de Letras.

Bibliografia: f. 151-159.

Apêndices: f. 160-206.

Exigências do sistema: Adobe Acrobat Reader.

1. Língua portuguesa – Estudo e ensino – Teses. 2. Língua portuguesa – Ortografia e silabação – Teses. 3. Língua portuguesa – Escrita – Teses. I. Guimarães, Daniela Mara Lima Oliveira. II. Universidade Federal de Minas Gerais. Faculdade de Letras. III. Título.

CDD : 469.07

Ficha catalográfica elaborada pelo Bibliotecário Israel José da Silva – CRB/6-2128
Biblioteca Professor Rubens Costa Romanelli - FALE/UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS
FACULDADE DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS LINGUÍSTICOS

FOLHA DE APROVAÇÃO

“CEBOLA” OU “SEBOLA”?: a evidência de múltiplos padrões no processo de escrita de grafemas concorrentes irregulares

ANA LUIZA DE SOUZA COUTO

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTUDOS LINGUÍSTICOS, como requisito para obtenção do grau de Doutor em ESTUDOS LINGUÍSTICOS, área de concentração LINGUÍSTICA APLICADA, linha de pesquisa Ensino do Português.

Aprovada em 03 de abril de 2024, pela banca constituída pelos membros:

Prof(a). Daniela Mara Lima Oliveira Guimarães - Orientadora
UFMG

Prof(a). Thais Cristofaro Alves da Silva
UFMG

Prof(a). Tatiana Cury Pollo
UFSJ

Prof(a). Adriane Teresinha Sartori
UFMG

Prof(a). Raquel Marcia Fontes Martins
UFLA

Belo Horizonte, 03 de abril de 2024.

Documento assinado eletronicamente por **Daniela Mara Lima Oliveira Guimaraes**, **Professora do Magistério Superior**, em 04/04/2024, às 16:43, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Raquel Marcia Fontes Martins**, **Usuário Externo**, em 04/04/2024, às 17:14, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Tatiana Cury Pollo**, **Usuário Externo**, em 04/04/2024, às 21:43, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Thais Cristofaro Alves da Silva**, **Professora do Magistério Superior**, em 04/04/2024, às 23:13, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Adriane Teresinha Sartori**, **Professora do Magistério Superior**, em 05/04/2024, às 07:59, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3090125** e o código CRC **7BA0FC7C**.

Referência: Processo nº 23072.213865/2024-71 SEI nº 3090125

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Dedico este trabalho aos meus pais, Teófila e Carlos, e ao meu irmão, Marcus. Vocês são minha fonte de força, meu porto seguro e minha inspiração constante.

AGRADECIMENTOS

Gostaria de agradecer a Deus, pois foi Ele quem me sustentou ao longo desta caminhada.

À minha orientadora, Daniela, que, com seu coração gigante, me guiou no caminho da pesquisa e sempre acreditou em meu potencial. Muito obrigada pela escuta, incentivo e parceria!

Agradeço aos meus pais, Aparecida e Betinho, que, desde sempre, me ensinaram que o conhecimento é algo que ninguém nos tira e que estudar é fundamental. Ao meu irmão, Marcus, pela parceria, cumplicidade e por sempre me apoiar e se interessar pelo mundo da pesquisa e da linguística. Ao meu primo, Carlos César, que sempre me apoiou e se alegra com cada conquista minha.

Sou grata à Adriane Sartori, à Raquel Martins, à Tatiana Pollo e à Thaïs Cristófaro pelas valiosas contribuições durante a banca. Um agradecimento especial à Tatiana e à Thaïs, que me acompanharam desde a qualificação.

À Vivi, minha roommate, por sempre estar ao meu lado e me ajudar em todos os momentos. Sua sabedoria e proatividade são inspiradoras e me ensinam muito.

À minha amiga Cissa, que prontamente me ajudou em diversas fases da pesquisa e me ajuda em outros aspectos da vida. Sua amizade e parceria têm tornado a vida acadêmica mais leve e significativa.

Ao meu amigo e parceiro de pesquisa, Marcelo, agradeço a colaboração e as valiosas discussões sobre o ensino da Língua Portuguesa. Sua seriedade e comprometimento com a pesquisa e com a sala de aula são admiráveis.

Aos meus amigos que caminham comigo nesta jornada acadêmica, Thais Bechir, Luiza Vignoli, Letícia Pena, Luiz, Lucas, Flora e Alcione, entre outros, obrigada pela companhia e apoio.

Agradeço também aos professores que contribuíram para o desenvolvimento desta pesquisa e para minha formação enquanto pesquisadora, Maria Cantoni, Jairo Carvalhais, Mahayana Godoy, Sueli Coelho, entre outros.

Aos integrantes do Grupo de Pesquisa sobre Práticas de Ensino de Escrita e Oralidade (PENSEO), obrigada pelas valiosas discussões sobre a Linguística Aplicada.

Por fim, agradeço à UFMG e ao POSLIN pela acolhida e por terem me apresentado um novo mundo. A universidade é para quem quiser!

“A arma social de luta mais poderosa é o domínio da linguagem”.

Magda Becker Soares

RESUMO

A norma ortográfica da Língua Portuguesa baseia-se no sistema alfabético de escrita, o qual consiste na correspondência entre fonemas e grafemas. Essa correspondência é constituída por múltiplas relações, que podem ser regulares diretas, regulares contextuais e irregulares (Lemle, [1994] 2009; Soares, 2018). A escrita ortográfica é uma das facetas da língua, a qual, atrelada a outros componentes e processos de aprendizagem, contribui para que o aluno seja alfabetizado e letrado, e que saiba fazer o uso da escrita nos mais diversos contextos sociais. No que diz respeito à ortografia nos documentos educacionais brasileiros, na Base Nacional Comum Curricular – BNNC (Brasil, 2017, 2018), espera-se que o aluno chegue ao final da Educação Básica, 3º ano do Ensino Médio, com a capacidade de fazer uso consciente da escrita ortográfica. No entanto, trabalhos apontam uma defasagem no domínio da ortografia por parte dos alunos ao longo dos anos escolares e após estes (Germani, 2017; Nascimento; Henz, 2020; Teis-Adamante; Busse, 2022). Além disso, as irregularidades ortográficas se configuram como uma das maiores dificuldades dos alunos no processo de aprendizagem da ortografia (Sartori; Mendes; Costa, 2015; Marquardt; Busse, 2015; Souza 2019; Castro, 2022). Neste contexto, o objetivo desta pesquisa é investigar a influência do ano escolar, da relação fonema-grafema e a frequência de tipo e de ocorrência na escrita de grafemas concorrentes irregulares de alunos de uma escola de Belo Horizonte. O arcabouço metodológico desta tese é pautado na abordagem mista quali-quantitativa (Paiva, 2019), do tipo de pesquisa de campo experimental (Lakatos; Marconi, 2003) e descritiva (Andrade, 2010). Esta pesquisa é guiada pela Teoria da Integração dos Múltiplos Padrões (IMP), proposta por Treiman e Kessler (2014). A IMP considera a aprendizagem da ortografia como múltipla e influenciada por diversos conhecimentos (Treiman; Kessler, 2014; Treiman; Decker; Kessler, 2017; Treiman, 2020). Os resultados indicam que o ano escolar, a relação fonema-grafema, a frequência de tipo e de ocorrência podem, em interação, influenciar a escrita de grafemas concorrentes irregulares. Portanto, esta pesquisa pode nortear a proposição de intervenções pedagógicas no ensino da ortografia que passem a considerar a relação fonema-grafema e a frequência no ensino de palavras de ortografia irregular.

Palavras-chaves: Escrita; Ortografia; Múltiplos padrões; Irregularidades ortográficas; Frequência de tipo; Frequência de ocorrência.

ABSTRACT

The orthographic standard of the Portuguese language is based on the alphabetic writing system, which consists of the correspondence between phonemes and graphemes. This correspondence is made up of multiple relationships, which can be direct regular, contextual regular and irregular (Lemle, [1994] 2009; Soares, 2018). Spelling is one of the facets of the language, which, linked to other components and learning processes, contributes to the student being literate, and knowing how to use writing in the most diverse social contexts. With regard to spelling in Brazilian educational documents, in the Base Nacional Comum Curricular – BNNC (Brazil, 2017, 2018), it is expected that the student reaches the end of basic education, 3ª Série do Ensino Médio, with the ability to do conscious use of spelling. However, studies indicate a lag in students' mastery of spelling throughout the school years and beyond (Germani, 2017; Nascimento; Henz, 2020; Teis-Adamante; Busse, 2022). Furthermore, orthograph irregularities are one of the biggest difficulties faced by students in the process of learning spelling (Sartori; Mendes; Costa, 2015; Marquardt; Busse, 2015; Souza 2019; Castro, 2022). In this context, the objective of this research is to investigate the influence of school year, the phoneme-grapheme relationship and the frequency of type and occurrence in the writing of irregular competing graphemes by students at a school in Belo Horizonte. The methodological framework of this thesis is based on a mixed qualitative-quantitative approach (Paiva, 2019), experimental field research (Lakatos; Marconi, 2003) and descriptive (Andrade, 2010). This research is guided by the Integration of Multiple Partners proposed by Treiman and Kessler (2014). The IMP considers spelling learning to be multiple and influenced by diverse knowledge (Treiman; Kessler, 2014; Treiman; Decker; Kessler, 2017; Treiman, 2020). The results indicate that grade level, the phoneme-grapheme relationship, type frequency, and token frequency may, in interaction, influence the spelling of irregular competing graphemes. Therefore, this research can guide the proposal of pedagogical interventions in the teaching of spelling that begin to consider the phoneme-grapheme relationship and the frequency in the teaching of words with irregular spelling.

Keywords: Writing; Spelling; Multiple partners; Orthographic irregularities; Type frequency; Token frequency.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ASPA	Projeto Avaliação Sonora do Português Atual
BNCC	Base Nacional Comum Curricular
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil
CEP	Comitê de Ética em Pesquisa
CCV	Consoante-consoante-vogal
CV	Consoante-vogal
CVC	Consoante-vogal-consoante
EF	Ensino Fundamental
EJA	Educação de Jovens e Adultos
EM	Ensino Médio
Enem	Exame Nacional do Ensino Médio
IMP	Integração dos Múltiplos Padrões
Inep	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
LDs	Livros Didáticos
LexPorBR	Léxico do Português Brasileiro
log 10	Logaritmo na base 10
MG	Minas Gerais
NILC	Núcleo Interinstitucional de Lingüística Computacional
Orto freq	Frequência da Ortografia
Orto freq/M	Frequência da Ortografia por milhão
PAF	Palavras de Alta Frequência
PB	Português Brasileiro
PBF	Palavras de Baixa Frequência
PCNs	Parâmetros Nacionais Educacionais
PENSEO	Grupo de Pesquisa sobre Práticas de Ensino de Escrita e Oralidade
PNLD	Programa Nacional do Livro Didático
p.	Página
TALE	Termo de Assentimento Livre e Esclarecido
TCLE	Termo de Consentimento Livre e Esclarecido
UFMG	Universidade Federal de Minas Gerais
V	Vogal
VOC	Vocabulário Ortográfico Comum da Língua Portuguesa

SUMÁRIO

1 INTRODUÇÃO

Qual a forma correta, de acordo com a ortografia do português brasileiro (doravante PB), de escrever: "cebola" ou "sebola"? Certamente, em algum momento de nossas vidas, nos deparamos com dúvidas sobre a forma escrita de determinadas palavras. Esse questionamento diz respeito ao que chamamos de ortografia. Cagliari e Massini-Cagliari (1999, p.98) afirmam que a função da ortografia é neutralizar variações linguísticas e "estabelecer um padrão de escrita que fosse único para todos os falantes, independentemente de como cada um fala o seu dialeto[1]". Morais (1998, p.18), por sua vez, define ortografia como "uma convenção social cuja finalidade é ajudar a comunicação escrita". Tais definições seguem o mesmo caminho atribuindo à ortografia uma importância na comunicação dos falantes por meio da escrita.

A ortografia da Língua Portuguesa baseia-se no sistema alfabético de escrita, o qual consiste na correspondência entre fonemas e grafemas. Essa correspondência é constituída por diversas relações, que podem ser biunívocas (um som para um grafema), regulares contextuais ou irregulares (Lemle, [1994] 2009; Soares, 2018). Segundo Nóbrega (2013, p.12), a convenção ortográfica é "um complexo sistema", o qual estabelece a função do grafema na palavra, regula o uso de prefixos e sufixos, determina as funções dos radicais, assim como o uso de desinências nominais e verbais, dentre outros. Tais relações ortográficas foram e são objetos de pesquisa de diversos trabalhos (Lemle, [1994] 2009; Cagliari, [1997] 2009; Morais, 1998; Oliveira, 2005; Nóbrega, 2013; Soares, 2018).

Dessa forma, em uma sociedade grafocêntrica em que a escrita desempenha papel central em vários contextos sociais, o domínio da escrita ortográfica se faz necessário para a construção de um sujeito letrado. Como definido por Soares (2018), a escrita ortográfica é uma das facetas da língua, a qual, atrelada a outros componentes e processos de aprendizagem, contribui para que o aluno seja alfabetizado e letrado, e que saiba fazer o uso da escrita nos mais diversos contextos sociais. Nesta linha de raciocínio, a ortografia pode atuar como um dos mecanismos de inserção social. Por exemplo, a falta de conhecimento da ortografia pode limitar as oportunidades de entrada no mercado de trabalho e o potencial para progressão na carreira (Apel; Henbest; Masterson, 2019; Henbest e*t al.* 2020). Além disso, a aprendizagem e o domínio da escrita, na qual inclui, também, a ortografia, são assegurados pela Lei nº 9.394 de 20 de dezembro de 1996 (Brasil, 1996), que estabelece as diretrizes e bases da educação

[1] Dialeto pode ser entendido como um conjunto de falantes que compartilham semelhantes características linguísticas. Por exemplo, dialeto mineiro, dialeto carioca (Cristófaro Silva, 2017).

nacional brasileira. Portanto, além de a aprendizagem da escrita ortográfica ser importante para a construção de um sujeito letrado, ela configura-se como um direito, assegurado por lei, a todo cidadão brasileiro.

Esta pesquisa é pautada na concepção de que o conhecimento ortográfico não envolve apenas escrever corretamente, mas revela também um conhecimento sobre a escrita, a combinação de letras e relações entre os sistemas ortográfico e fonológico. Como observam Oliveira, Castro e Couto (2023, p.8), ao escrever 'essessão', "podemos dizer que utilizamos um conhecimento ortográfico", visto que <ss> entre vogais representa o som de [s]. Dessa forma, tendo tal linha de raciocínio como referência, o erro ortográfico, nesta pesquisa, é assumido como "elemento revelador do processo de aprender", o qual se trata de "um processo de análise e reorganização, e pode auxiliar aqueles que estudam a aquisição da linguagem a investigar o saber construído" (Miranda, 2010, p.4).

Com a ortografia como um dos enfoques desta pesquisa, o objetivo geral da tese é investigar a influência do ano escolar, do padrão ortográfico e a frequência de tipo e de ocorrência na escrita de grafemas concorrentes irregulares de alunos de uma escola de Belo Horizonte – MG. Os alunos participantes da pesquisa estão em diferentes anos escolares da Educação Básica, 3º ano do Ensino Fundamental I (EFI), 6º e 9º anos do Ensino Fundamental II (EFII) e 3ª série do Ensino Médio (EM). Para que o objetivo fosse alcançado, foram consideradas as seguintes relações irregulares dos fonemas /s/, /ʒ/ e /z/ e os grafemas que os representam, sistematizadas no Quadro 1 a seguir.

Quadro 1 - Relações ortográfica irregulares pesquisadas nesta tese

Fonema	Grafema	Contexto	Exemplo
/s/	<c> ou <s>	Em início de palavra, diante de <i> ou <e>	**c**inema e **s**igilo **s**eco e **c**ego
/ʒ/	<g> ou <j>	Diante de <i> ou <e> em início de palavra	**g**igante e **j**iboia **g**eleia e **j**erico
/z/	<z> ou <s>	Entre vogais	natur**z**a e me**s**a a**z**edo e fra**s**e

Fonte: Elaborado a partir de Soares (2018).

A partir da leitura do Quadro 1, pode-se entender que os grafemas <c> e <s>, no contexto inicial de palavra diante de <i> e <e>, concorrem para representar o mesmo fonema /s/. De modo semelhante, o fonema / ʒ/, diante de <i> e <e>, pode ser representado por dois

grafemas, o <g> e o <j>. Por fim, os grafemas <z> e <s> concorrem para representar o fonema /z/ em contexto intervocálico. Em linhas gerais, a frequência tem relação com o número de vezes em que determinado evento ocorre (Cristófaro Silva, 2017). Nesta pesquisa, a frequência de ocorrência é o número de vezes que uma palavra, com um dos grafemas dos pares do Quadro 1, já a frequência de tipo tem relação a quantas vezes os grafemas pesquisados ocorrem. Na seção seguinte, as justificativas da pesquisa são elucidadas.

1.1 Relevância científica

Santaella (2001) afirma que as justificativas de uma pesquisa científica podem ser de ordem social, científica-prática e científica-teórica. Na ordem social, esta pesquisa se justifica, pois, ao investigar a escrita ortográfica de alunos da educação pública. No Brasil, de acordo com a Sinopse Estatística da Educação Básica 2022 (Brasil, 2023), 86,05 % (mais de 34,3 milhões) dos alunos matriculados na Educação Básica brasileira são de escolas públicas, enquanto apenas 13,95% (5,5 milhões) são de instituições privada. Neste contexto, desenvolver esta pesquisa em escolas públicas se justifica pelo intuito de contribuir para a formação dos alunos a fim de que estes se tornem sujeitos letrados capazes de utilizar a escrita nos mais diversos contextos sociais.

Na perspectiva científica-prática, esta pesquisa se justifica ao observar um desencontro entre os documentos educacionais, como a Base Nacional Comum Curricular – BNCC (Brasil, 2017, 2018), e a realidade da escrita ortográfica de alunos brasileiros. De acordo com a BNCC (Brasil, 2017, 2018), o aluno, ao final do Ensino Fundamental I, precisa escrever palavras seguindo as correspondências regulares, contextuais e morfológicas, além de conseguir grafar palavras irregulares de uso frequente. Ao final do Ensino Fundamental II, o discente, segundo a Base, precisa ter domínio da escrita ortográfica e, no Ensino Médio (último segmento da Educação Básica brasileira), fazer uso consciente da ortografia padrão. No entanto, pesquisas revelam uma discrepância entre a realidade da escrita ortográfica de alunos da Educação Básica brasileira e o que se espera nos documentos educacionais. Há problemas na escrita ortográfica de alunos em todos os segmentos educacionais, desde o Ensino Fundamental I (Miranda, 2012, 2020; Batista; Capellini, 2017), Ensino Fundamental II (Freitas, 2011; Souza, 2020; Oliveira; Castro; Couto, 2023) até o Ensino Médio (Sartori; Mendes; Costa, 2015; Selle, 2017; Oliveira, 2021; Lacerda; Couto; Oliveira, 2023).

Dentre as principais dificuldades com a escrita ortográfica, destaque é dado às irregularidades ortográficas, pois são relações que quaisquer usuários da escrita do PB,

inclusive alunos que estão em processo de aprendizado, têm dificuldades em compreendê-las e em grafá-las (Sartori; Mendes; Costa, 2015; Nobile; Barrera, 2016; Marquardt; Busse, 2015; Saggiomo, 2018; Souza 2019; Castro, 2022). Nesse contexto, investigar a influência do ano escolar, do padrão ortográfico e da frequência na escrita de grafemas concorrentes irregulares de alunos de escolas públicas torna-se imprescindível para a compreensão de uma das maiores dificuldades dos discentes em relação à escrita ortográfica.

Por fim, sob a perspectiva de ordem científica-teórica, esta tese se justifica ao propor uma investigação sobre o conhecimento ortográfico dos alunos em relação às irregularidades ortográficas. Esta pesquisa avança, em relação à literatura precedente, ao investigar diferentes fatores no processo de escrita de grafemas concorrentes irregulares, como o ano escolar e a frequência de tipo e de ocorrência. Até o desenvolvimento desta pesquisa, não foram encontrados trabalhos desenvolvidos sobre este tema em relação ao português brasileiro. Por fim, esta tese se justifica, pois visa contribuir com avanços para a Teoria da Integração dos Múltiplos Padrões – IMP (Treiman; Kessler, 2014) na perspectiva do português brasileiro (PB).

1.2 Organização teórica

Esta tese se concentra no campo da Linguística Aplicada, que abarca a "teorização em que teoria e prática sejam conjuntamente consideradas em uma formulação do conhecimento" (Lopes, 2008. p.101). O percurso metodológico deste trabalho é pautado na abordagem mista quali-quantitativa (Paiva, 2019), do tipo de pesquisa de campo experimental (Lakatos; Marconi, 2003) e descritiva (Andrade, 2010).

A teoria IMP (Treiman; Kessler, 2014; Pollo; Treiman; Kessler, 2015; Treiman, 2017; Treiman e*t al*., 2019; Treiman; Kessler, 2022; Oliveira; Castro; Couto, 2023) é o arcabouço teórico e guia desta tese, a qual busca verificar a hipótese de que a escrita ortográfica de grafemas concorrentes irregulares é influenciada por múltiplos fatores, como o ano escolar, a relação fonema-grafema, a frequência de tipo e de ocorrência. Além desta introdução, esta tese está organizada nos seguintes Capítulos:

- Capítulo 2 - Panorama da Pesquisa: objeto de estudo, hipótese e objetivos
    - Objetivo do Capítulo: descrever o objeto de estudo desta tese e apresentar as questões de pesquisa, os objetivos que norteiam este trabalho e a hipótese defendida.
- Capítulo 3 - Estado da Arte: caminhos percorridos pela e na ortografia

    - Objetivos do Capítulo: desenvolver uma discussão sobre os caminhos percorridos pela e na ortografia do português brasileiro; descrever a ortografia do PB e discutir a organização das relações ortográficas; refletir sobre o lugar da ortografia na Base Nacional Comum Curricular – BNCC (Brasil, 2018). Por fim, discutir trabalhos científicos sobre a ortografia de alunos da Educação Básica brasileira e do Ensino Superior.
- Capítulo 4 - A Frequência em Evidência
    - Objetivo do Capítulo: discutir sobre o papel da frequência na Linguística, especificamente à aprendizagem da ortografia.
- Capítulo 5 - Arcabouço Teórico
    - Objetivos do Capítulo: situar a tese no campo da Linguística Aplicada e apresentar a teoria guia desta pesquisa, a IMP (Treiman; Kessler, 2014).
- Capítulo 6 - Percursos Metodológicos
    - Objetivos do Capítulo: caracterizar a abordagem metodológica desta tese, além de descrever os caminhos metodológicos percorridos.
- Capítulo 7 - Análise de Dados e Discussão dos Resultados
    - Objetivos do Capítulo: analisar, quali-quantitativamente, os dados e discutir sobre os resultados.
- Capítulo 8 - Considerações Finais
    - Objetivos do Capítulo: apresentar as considerações finais e a prospecção de futuras pesquisas.

2 PANOROMA DA PESQUISA: OBJETO DE ESTUDO, HIPÓTESE E OBJETIVOS

O objetivo deste Capítulo é descrever o objeto de estudo desta tese, assim como apresentar as questões de pesquisa que norteiam este trabalho e a hipótese a ser defendida. Além disso, serão apresentados objetivo geral e os objetivos específicos.

2.1 Objeto de estudo

O objeto de estudo desta tese é a escrita de grafemas concorrentes irregulares, ou seja, a escrita de relações ortográficas irregulares. Por isso, o conhecimento ortográfico dos alunos, participantes deste trabalho é importante. Portanto, nesta pesquisa, o conhecimento ortográfico é compreendido como o conjunto de informações que o aluno possui sobre como escrever determinada palavra. Ou seja, até mesmo quando um aluno erra a ortografia de alguma palavra, ele está mobilizando um conhecimento ortográfico (Oliveira, Castro e Couto, 2023). Portanto, o objeto de estudo desta tese é o conhecimento ortográfico de alunos da Educação Básica em relação às irregularidades ortográficas.

Ora, o que seria uma irregularidade ortográfica? Na ortografia do PB, há indícios de regularidades na escrita ortográfica de determinadas palavras. Por exemplo, o som [p] sempre será representado pela letra <p>, como [ˈpatʊ] escrito como “pato”. Por outro lado, há palavras que dependem do contexto para regular a escrita de determinada palavra. Por exemplo, o grafema <m>, em posição final de sílaba, é escrito diante de <p> e <b>, como em “pomba”. Por fim, há casos em que não há uma relação direta entre fonema e grafema, e nem o contexto pode regular determinada escrita ortográfica. Nestes casos, dois ou mais grafemas podem concorrer para representar o mesmo fonema. No caso dos grafemas <c> e <s>, que ocorrem para representar o fonema /s/, há palavras como “seco” e “cego”.

Há evidências na literatura de que as relações ortográficas irregulares são aquelas em que os alunos têm uma notável dificuldade de domínio durante o processo de aprendizagem da ortografia (Sartori; Mendes; Costa, 2015; Marquardt; Busse, 2015; Nobile; Barrera, 2016; Souza 2019; Castro, 2022). Mesmo que essas relações sejam o foco de abordagem em livros didáticos, como apontam as pesquisas de Teis-Adamante e Parise (2018) e Couto (2020), elas apresentam um empecilho na aprendizagem da ortografia e levam a um maior número de erros ortográficos nas palavras com tais irregularidade (Souza, 2019; Castro, 2022). Neste contexto,

é importante investigar o conhecimento ortográfico dos alunos em relação a essas irregularidades.

Para ter acesso a este conhecimento ortográfico dos alunos em relação às irregularidades ortográficas, focamos entre três casos: a escrita de <c> ou <s> para representar o fonema /s/, <g> ou <j> para representar o /ʒ/ e, por fim, <z> ou <s> para representar o fonema /z/. Por exemplo, o fonema /s/ pode ser escrito pelos grafemas <c> ou <s> no mesmo contexto, diante de <i> e <e> em início de palavra, como "cinema" e "sigilo", "seco" e "cego". Isso quer dizer que esses dois grafemas competem entre si para representar o mesmo fonema em determinadas situações. De modo semelhante, o fonema /ʒ/ pode ser representado por dois grafemas (<g> e <j>) em um mesmo contexto, quando diante de <i> e <e>, como "gigante" e "jiboia", "geleia" e "jerico". Vale ressaltar que, nesta tese, optamos por restringir o contexto do uso de <g> e <j> também no início de palavras, visto que o contexto desempenha um papel é importante na aprendizagem da escrita pelos alunos (Treiman; Kessler, 2014; Toledo, 2023). No entanto, sabemos que a concorrência do <g> e <j> diante de <i> ou <e> pode ocorrer em qualquer parte da palavra, como em "sujeito". Por fim, observa-se que os grafemas <z> e <s> concorrem para representar o fonema /z/ quando estão entre vogais, como nas palavras "azedo" e "frase". Embora haja alguns casos, no português brasileiro, de palavras escritas com <x>, mas com som de [z] nesta tese, o foco é dado aos grafemas <s> e <z> entre vogais que representam o fonema /z/, por considerarmos o contexto como um elemento importante na linguística, como evidenciado no trabalho de Toledo (2023).

Em uma pesquisa experimental, Toledo (2023) descobriu que as crianças, no processo de aprendizagem da escrita ortográfica, rapidamente compreendem que as vogais átonas finais [ɪ], [ʊ], por exemplo, são, categoricamente, escritas como <e> e <o>. Por exemplo, palavras pronunciadas como [ˈpẽtʃɪ] e [ˈovʊ] são escritas ortograficamente como "pente", "ovo". Por outro lado, os mesmos sons [ɪ] e [ʊ] em um contexto diferente, o pretônico, são aprendidos tardiamente pelos alunos. A autora identificou erros em palavras como "perigo", "pirata", onde casos como *perata[2] e *pirigo foram encontrados pela autora. Os resultados indicaram que as trocas de <e, o> por <i, u> na escrita persistem em diferentes anos escolares a depender do contexto acentual, pretônico ou postônico final.

Além de Toledo (2023), outros trabalhos ressaltam a importância do contexto no processo de aprendizagem da escrita (Treiman, Kessler, 2014). De acordo com a IMP – arcabouço teórico desta tese – (Treiman; Kessler, 2014), o aprendiz que considera o contexto

[2] Nesta tese, o asterisco é utilizado antes de palavras que foram escritas em desacordo com a ortografia.

em seu processo de aprendizagem tem mais chance de escrever de acordo com a ortografia. Portanto, nesta tese, consideramos o contexto como um importante elemento no processo de aprendizagem da escrita ortográfica. Com o contexto como elemento importante no processo de aprendizagem, optamos por trabalhar com dois contextos: os fonemas /s/ e /ʒ/ em início de palavra, diante de <i> e <e>, e o fonema /z/ entre vogais. Portanto, esta tese se propõe a investigar a escrita de três casos de irregularidade ortográfica em dois contextos específicos: os fonemas /s/ e /ʒ/ em posição inicial de palavra, diante das vogais <i> e <e>; e o fonema /z/ entre vogais. Neste contexto linguístico, serão a frequência de ocorrência de palavras com esses grafemas e a frequência de tipo, ou seja, quantas vezes estes grafemas ocorrem.

2.2 Questão e Hipótese de Pesquisa

A partir da discussão realizada anteriormente, pode-se inferir que aprender a ortografia pode estar relacionado a múltiplos padrões, como as relações regulares e irregulares e aos diversos padrões ortográficos, que podem envolver a sílaba, o grafema, o fonema e o contexto de ocorrência do grafema na palavra. Além disso, vale ressaltar que a frequência tem relação com o desempenho ortográfico dos alunos (Pacton *et al.*, 2001; Santos; Befi-Lopes, 2013; Nigro *et al.*, 2014; Fay; Hein; Ghayoomi, 2015; Ribeiro; Martins, 2020). A frequência parece ser um elemento importante no processo de aprendizagem da ortografia. Deste contexto de dificuldades com a escrita de grafemas concorrentes, atrelado à importância da frequência no processo de aprendizagem, surge o objeto de investigação desta pesquisa. Além disso, considera-se, nesta pesquisa, o ano escolar como um importante aspecto de investigação, visto que os erros ortográficos são frequentes ao longo dos anos escolares da Educação Básica brasileira, desde o Ensino Fundamental I (Miranda, 2012, 2020; Batista; Capellini, 2017), no Ensino Fundamental II (Souza, 2020; Castro, 2022) até o Ensino Médio (Sartori; Mendes; Costa, 2015; Selle, 2017; Oliveira *et al.*, 2021; Lacerda; Couto; Oliveira, 2023).

Neste contexto, origina-se a seguinte questão norteadora desta pesquisa*: O ano escolar, a relação de determinados fonemas e seus grafemas, a frequência de tipo e de ocorrência podem influenciar a escrita ortográfica de grafemas caracterizados como concorrentes irregulares?*

Para responder a esta pergunta esta tese busca defender a seguinte hipótese: A escrita ortográfica de grafemas concorrentes irregulares é influenciada por múltiplos fatores, como (a) o ano escolar, (b) a relação fonema-grafema, (c) a frequência de tipo e (d) a frequência de ocorrência. Esta hipótese está organizada no quadro a seguir:

Quadro 2 - Quadro de hipótese da pesquisa.

Hipótese	
A escrita ortográfica de grafemas concorrentes irregulares é influenciada por múltiplos fatores, como	(a) o ano escolar
	(b) a relação fonema-grafema
	(c) frequência de tipo
	(d) frequência de ocorrência

Fonte: Elaboração própria.

Espera-se que os resultados desta pesquisa possam confirmar que os fatores (a), (b), (c) e (d) influenciam, de forma combinada, a escrita ortográfica dos grafemas concorrentes irregulares.

2.3 Objetivos da Pesquisa

O objetivo geral desta pesquisa é investigar a influência do ano escolar, do padrão ortográfico e a frequência de tipo e de ocorrência na escrita de grafemas concorrentes irregulares de alunos de uma escola de Belo Horizonte – MG. Para que o objetivo geral seja alcançado, os seguintes objetivos específicos foram elaborados:

- descrever os padrões ortográficos do português brasileiro relacionados aos pares de grafemas concorrentes irregulares <c> e <s>, <g> e <j>, <z> e <s>;
- mapear a frequência de tipo envolvendo os pares de grafemas concorrentes irregulares <c> e <s>, <j> e <g> e <z> e <s>;
- mapear a frequência de ocorrência de palavras com os grafemas concorrentes irregulares <c> e <s>, <j> e <g> e <z> e <s>;
- relacionar a escrita de grafemas concorrentes irregulares dos participantes da pesquisa ao ano escolar;
- relacionar a escrita de grafemas concorrentes irregulares ao tipo do padrão ortográfico;
- investigar a relação entre a escrita de grafemas concorrentes irregulares e a frequência de tipo;
- investigar a relação entre a escrita de grafemas concorrentes irregulares e a frequência de ocorrência.

3 ESTADO DA ARTE: CAMINHOS PERCORRIDOS PELA E NA ORTOGRAFIA

O objetivo deste Capítulo é abordar os caminhos percorridos pela e na ortografia do português brasileiro. Para isso, na primeira seção, descrevemos a ortografia do PB e refletimos sobre a sua categorização, destacando os teóricos que propuseram a organização das relações ortográficas. Além disso, na segunda seção, discutimos sobre o lugar da ortografia no importante documento educacional brasileiro, a Base Nacional Comum Curricular – BNCC – (Brasil, 2017, 2018). Por fim, na última seção, apresentamos, por meio de trabalhos científicos, a escrita ortográfica de alunos da Educação Básica brasileira e de discentes do Ensino Superior.

Para aprimorar o entendimento deste Capítulo, é importante compreender o que é fonema, fone, letra, grafema e padrão ortográfico. Estes termos, essenciais para compreender a argumentação desta tese, estão organizados no quadro abaixo, juntamente às suas definições e exemplos.

Quadro 3 - Nomenclaturas importantes para esta tese

Nomenclatura	Definição	Exemplo
Fonema	Unidade abstrata sonora, vocálica ou consonantal, capaz de distinguir palavras em uma língua. Fonema é transcrito em barras transversais /a/	/f/ →/ˈfa.ka/ /v/ →/ˈva.ka/
Fone	Unidade sonora vocálica ou consonantal que possui correlatos acústicos. Fone é sempre transcrito entre colchetes [a]	[f] → [ˈfa.kə] [v] → [ˈva.kə]
Grafema	Unidade abstrata do sistema gráfico. O grafema é sempre transcrito como <a>	<a>, <b>, <ch>
Letra	Unidade concreta do sistema gráfico. A letra pode variar de tamanho, cor, estilo	A, ***a***, 𝐀, 𝖆, a B, ***b***, 𝐁, 𝖇, b
Padrão ortográfico	Um ou mais semelhanças na escrita de determinada língua, como o fonema e o grafema. Usa-se, nesta tese, o sinal <a> para representá-lo	<o> → beijo, queijo, bolo; <ci> → cidade, cilada

Fonte: Elaborado a partir de Cristófaro Silva (2017), Carvalho (2014) e Treiman e Kessler (2014).

O fonema é uma unidade abstrata e objeto de estudo da Fonologia e, como observado no quadro anterior, uma de suas funções é a distinção de palavras quanto ao significado. Já o fone é uma unidade concreta, objeto de estudo da Fonética (Cristófaro Silva, 2017).

O grafema, por sua vez, é a unidade abstrata do sistema de escrita (Carvalho, 2014), o qual também atua na distinção de palavras, como <v> em "vaca" e <f> em "faca". Além disso, o grafema pode ser composto por dois itens, como o <ch>, comumente denominado de dígrafo. Já a letra é um "termo genérico", que possui um caráter particular ao indivíduo, visto que ela

pode modificar a depender do estilo de escrita, da caligrafia de quem escreve (Carvalho, 2014). As variações da letra A, ***a***, 𝙖, 𝖺, a representam uma mesma unidade abstrata, o grafema <a>.

Por fim, o padrão ortográfico, também chamado de *spelling patterns*, trata de uma ou mais semelhanças que são compartilhadas pelas palavras em determinada língua (Treiman; Kessler, 2014). Por exemplo, no português brasileiro, as vogais átonas finais reduzidas, como em "beijo" [ˈbe.ʒʊ] e "pente" [ˈpẽ.tʃɪ], são grafadas, categoricamente, como <o> e <e> na escrita. Esta pesquisa adota o conceito de grafema e fonema, por serem elementos abstratos que compõem a representação abstrata dos usuários de uma língua. Já o termo letra será utilizado para referir ao que o participante da pesquisa escreveu.

3.1 Reflexões sobre a ortografia do português brasileiro

O objetivo desta seção é desenvolver uma reflexão sobre a organização da ortografia do PB e discutir sobre as suas relações ortográficas. Para compreender a ortografia, é necessário, primeiramente, a contextualização do sistema de escrita da Língua Portuguesa e da função da ortografia na língua e na sociedade.

A Língua Portuguesa, língua românica originada do Latim, utiliza o sistema da escrita alfabética. Este sistema representa os sons da língua por meio de sinais gráficos (Massini-Cagliari; Cagliari, 1999). Em outras palavras, os fonemas da língua são representados pelos grafemas no sistema de escrita alfabética. Até um certo momento, apenas a escrita alfabética supriu as demandas da sociedade em relação à comunicação pela escrita. No entanto, por não haver uma regulamentação para esta escrita, dificuldades para compreender o que foi escrito começaram a surgir. Neste contexto, surgiu a ortografia.

Segundo Cagliari (1999), se não fosse a ortografia, a escrita alfabética não teria dado certo, visto que, com a escrita alfabética sem a ortografia, a variação linguística de quem escreve seria grafada. Por exemplo, um indivíduo de Belo Horizonte – MG escreveria a palavra "beijo" como ele fala, "beju" [ˈbe.ʒʊ]. Neste contexto, a ortografia surgiu como uma "tecnologia" (Morais; Teberosky, 1994), justamente para neutralizar a variação sonora na escrita (Cagliari, [1997] 2009). Portanto, a ortografia é "uma convenção social cuja finalidade é ajudar a comunicação escrita" facilitando-a (Morais, 1998, p.18). Além dos grafemas, há, na escrita ortográfica, os "sinais diacríticos", tais como acento agudo, acento grave e circunflexo e o til (Cagliari, [1997] 2009). Nesta tese, destaque é dado apenas às relações entre grafemas e fonemas, e não aos diacríticos.

A ortografia do português passou por constantes transformações até chegar à forma que conhecemos hoje. Em linhas gerias, a história da ortografia do português pode ser classificada de acordo com três períodos, o fonético, o pseudoetimológico e o simplificado (Coutinho, 1986; Neves, 2010). No primeiro período, denominando fonético, também conhecido como a fase arcaica do português, que ocorreu por volta do século XVI, houve uma tentativa de escrever de acordo com a fala, pois "escrevia-se não para as vistas, mas para o ouvido" (Coutinho, 1986, p.72). A comunicação pela escrita, neste primeiro período, passou a enfrentar barreiras, visto que pessoas de diferentes dialetos na mesma língua não se entendiam.

Com o aumento do interesse pelo Latim, e a valorização dos estudos clássicos no Renascimento, por volta do século XVI ao XX, iniciou-se o segundo período, o pseudoetimológico. Segundo Coutinho (1984, p.75), o objetivo deste período foi de "respeitar, tanto quanto possível, as letras originárias da palavra, embora nenhum valor fonético". Com a inserção de novas letras e novas palavras na escrita, surgiram dificuldades em relação à leitura e a escrita, e, mais uma vez, a comunicação foi prejudicada. Além disso, no período pseudoetimológico, surgiram alguns indícios de tratados ortográficos, mas havia arbitrariedade no registro escrito. A partir do século XX até à atualidade, o período simplificado se instaurou diante da necessidade de elaborar um acordo ortográfico com o objetivo de simplificar e unificar a ortografia e, consequentemente, facilitar a comunicação por meio da escrita.

A ortografia da Língua Portuguesa passou por diversos acordos. A primeira reforma ortográfica foi 1911 desenvolvida apenas em Portugal. Posteriormente, a Academia Brasileira de Letras e Academia de Ciências de Lisboa uniram esforços em busca de uma ortografia unificada. Após diversos acordos, modificações e alterações, em 1990 foi divulgado o Novo Acordo Ortográfico da Língua Portuguesa, extensivo a todos os países lusófonos[3], como Brasil, Portugal, Timor Leste, Cabo Verde, Moçambique, Angola, Guiné-Bissau e São Tomé e Príncipe. No Brasil, o Acordo de 1990 entrou em vigor apenas em 2008, por meio do Decreto de Nº 6.583, foi promulgado o "Acordo Ortográfico da Língua Portuguesa, assinado em Lisboa, em 16 de dezembro de 1990" (Brasil, 2008).

Além de ser uma tecnologia, como bem definiu Morais e Teberosky (1994), a ortografia que hoje conhecemos passou por diversas mudanças motivadas por questões sociais, como a necessidade de comunicação, e, também, por questões jurídicas, para facilitar a "comunicação diplomática entre países lusófonos" (Neves, 2010, p.18). O acordo ortográfico, para os países

[3] [3] Países falantes da língua portuguesa.

envolvidos, “constitui um passo importante para a defesa da unidade essencial da língua portuguesa e para o seu prestígio internacional” (Brasil, 2008, p.1).

Dessa forma, para que um usuário do português consiga fazer uso da escrita em diversos contextos sociais, a ortografia, atrelada a outros aspectos da escrita, como os semânticos, textuais, sintáticos, morfológicos etc., se faz de extrema importância. Soares (2018) traz a discussão de que a ortografia também é uma importante faceta linguística. No entanto, infelizmente, a ortografia ainda é vista, por vezes, como conteúdo mecânico, que não faz parte da linguística, cujo único caminho para a sua aprendizagem é a memorização. O ensino da ortografia, por vezes, não está presente nas aulas de Língua Portuguesa da Educação Básica, ou caso o conteúdo ortográfico apareça é por meio de atividades mecânicas ou apenas pela correção dos textos dos alunos (Oliveira, Castro e Couto, 2023).

Entretanto, mesmo a ortografia não tendo um lugar consolidado na sala de aula, o seu domínio, atrelado a outros fatores, é cobrado, como em oportunidades de entrada no mercado de trabalho e o potencial para progressão na carreira (Apel; Henbest; Masterson, 2019; Henbest e*t al.* 2020), assim como na elaboração da redação do Exame Nacional do Ensino Médio (Enem), exame brasileiro que avalia o aluno para o ingresso em instituições de Ensino Superior no Brasil e em outros países. Portanto, a ortografia possui duas funções, a função linguística de cristalizar as variações da língua falada e a função social de levar o seu usuário a usá-la em diversos contextos sociais.

Pela base alfabética do PB, as relações ortográficas são estabelecidas a partir da relação entre fonema e grafema. Antes de adentrar às relações ortográficas, é importante ressaltar que o termo “regularidade ortográfica”, utilizado neste Capítulo, refere-se à relação entre fonema e grafema, e não ao termo regular definido pela frequência dos grafemas (Chetail, 2015). Um grafema pode ser considerado regular quando analisado a partir do critério da frequência de tipo. No entanto, ao considerar a relação fonema-grafema, ele pode ser uma irregularidade ortográfica por ter uma relação indireta com o fonema.

Por exemplo, o fonema /s/ pode ser representado por nove grafemas diferentes, <s> e <c> em início de palavras diante de <e> e <i>, <s> e <c> em outros contextos, <ss>, <ç>, <x>, <xc>, <sc>. De acordo com o Vocabulário Ortográfico Comum da Língua Portuguesa –VOC– (Ferreira *et al.*, 2017), o grafema <ss> ocorre 111.334 vezes, enquanto o grafema <xc> aparece 2.154 vezes. Dessa forma, com base na frequência, <ss> seria considerado regular por ter um alto número de ocorrências, enquanto <xc> seria irregular devido à sua baixa frequência. No entanto, pela relação fonema-grafema, ambos grafemas são irregulares, pois fazem parte de um

conjunto de possibilidades para representar o fonema <s>. Portanto, nesta tese, o termo "regularidade ortográfica" se refere à correspondência entre fonema e grafema.

Em estudos, como os de Lemle ([1994] 2009) e Morais (1998), se debruçaram na organização das relações ortográficas do português brasileiro. Soares (2018) retoma a discussão sobre as relações ortográficas na perspectiva da alfabetização, destacando a importância de conhecer e ter consciência da ortografia do PB para um satisfatório domínio da escrita. No quadro a seguir, há a descrição resumida sobre as relações ortográficas do PB.

Quadro 4 – Relações ortográficas do PB.

Relação ortográfica	Definição	Exemplo
(1) Regularidade direta	O mesmo fonema é representado pelo mesmo grafema, vice-versa	/p/ → <p> **p**ato /b/ → <b> **b**ato
(2) Regularidade contextual	O contexto regula o uso do grafema	/k/ → antes de a, o, u → <c> **c**avalo, sa**c**ola → antes de e, i → <qu> **qu**eda, es**qu**ina /u/ → em contexto átono final → <o> ov**o**
(3) Irregularidade	Relação arbitrária em que um mesmo fonema pode ser representado por diferentes grafemas	/s/ → <s> **s**ituação <c> **c**ebola /z/ → <s> ca**s**a <z> a**z**ar <x> e**x**ame /ʃ/ → <ch> **ch**inelo <x> **x**ícara /ʒ/ → <g> **g**ibi <j> **j**iló

Fonte: Elaborado a partir de Lemle ([1994] 2009) e Soares (2018).

Como se pode observar no quadro acima, cada grafema do alfabeto do PB possui a sua classificação, podendo ser regular, regular-contextual ou irregular, relacionada ao contexto de uso. A relação (1) regular indica que um mesmo fonema é representado pelo grafema, vice-versa. Como apontado por Lemle ([1994] 2009), há um relacionamento monogâmico entre som e grafema. Já na relação (2) regular contextual, o contexto define o uso de determinado grafema. Por fim, a relação (3), irregular, refere-se à concorrência entre grafemas para preencher determinada posição da palavra. Por exemplo, o fonema /s/ pode ser representado por sete grafemas, tais como <s>, <c>, <ss>, <sc>, <sç>, <x>, <xc>, sem que haja uma regra para a

escolha de qual deles escrever. Além do fonema /s/, há os fonemas /z/, /ʃ/ e /ʒ/ que também estabelecem uma relação irregular com os grafemas que os representam. Pelo modo de articulação, os sons [s], [z], [ʃ] e [ʒ] são chamados de fricativas, pois o modo de articulação destas consoantes ocorre por meio de uma fricção dos articulares (Cristófaro Silva, 2017). Nesta tese, o foco é dado a três casos de irregularidades ortográficas, a saber: o fonema /s/ e os grafemas <s> e <c>; o /ʒ/ e os grafemas <g> e <j>; o fonema /z/ e os grafemas <s> e <z>, todos os três casos em contexto inicial de palavras diante de <e> e <i>.

Ao propor uma organização para o ensino da ortografia, Morais (1998) estabeleceu a Regularidade morfológico-gramatical. Este pode ser um caminho para propor uma regularidade às irregularidades na relação fonema-grafema. Para o autor, a categoria gramatical pode estabelecer a regularidade ortográfica. Por exemplo, a terminação de adjetivos pátrios sempre será com <esa> (“francesa”); e o final de substantivos derivados de adjetivos sempre será <eza> (“pobreza”); e o uso de <r> em final de todos os verbos no infinitivo (“cantar” e “falar”). O conhecimento morfológico pode ajudar o aluno a resolver problemas de escrita ortográfica (Guimarães; Mota, 2018). No entanto, este conhecimento só é adquirido em anos escolares mais avançados. Como, nesta tese, os participantes são alunos de diversos anos escolares, e não medimos o conhecimento deles em relação à morfologia, optamos por considerar a relação entre /z/ e os grafemas <s> e <z> em contextos não regulares.

Ademais, vale ressaltar que, em uma mesma palavra, pode haver relações ortográficas regulares e irregulares. Por exemplo, na palavra “cidade”, o uso de <c> com som de [s], em início de palavras, diante de [e] e [i], é uma irregularidade, visto que, para esta mesma posição, há a concorrência entre o uso do grafema <c> ou <s>. Em outras palavras, não há nenhuma regra que regule o uso de <c> em “cidade”. Já o grafema <e> no final de “cidade” refere-se a uma regularidade contextual, visto que o som de [i], em contexto final de sílaba átona, é, categoricamente, representado pelo grafema <e>.

Como visto, na ortografia do PB, há relações diretas entre fonema e grafemas, mas também há irregularidades. Em outros idiomas, como finlandês, grego e italiano, há, também, diversos tipos de relações entre fonemas e grafemas. Seymour e colaborados (2003), em uma investigação sobre a aprendizagem inicial da leitura em 13 diferentes línguas, chegaram à conclusão de que a complexidade silábica e a profundidade ortográfica impactam no processo de aprendizagem da leitura. Os autores definem a complexidade silábica como o tipo de estrutura que é mais comum em uma língua, como a sílaba aberta, CV (consoante-vogal), ou a fechada CVC (consoante-vogal-consoante). O foco de tal trabalho foi na leitura, no entanto, podemos trazer esta reflexão para a aprendizagem da escrita.

Hipoteticamente, Seymour e colaborados (2003) classificaram a ortografia das 13 línguas quanto à profundidade ortográfica e à estrutura silábica. No quadro a seguir, há uma sistematização em relação à profundidade da ortografia destas línguas.

Quadro 5 - Profundidade ortográfica.

		Profundidade ortográfica				
		Transparente <----------	Relativamente transparente		Opaca ---------->	
Estrutura silábica	**Simples**	Finlandês	Grego Italiano Espanhol	Português Europeu	Francês	-
	Complexa		Alemão Norueguês Islandês	Holandês Sueco	Dinamarquês	Inglês

Fonte: Elaborado a partir de Seymour *et al.* (2003)

No Quadro 5, a coluna de estrutura silábica classifica as línguas em simples ou complexas. Quanto à profundidade ortográfica, as línguas são classificadas no continuum de opaca à transparente. Por exemplo, o inglês é de ortografia opaca e de estrutura silábica complexa. O finlandês, por sua vez, é de estrutura silábica simples e de ortografia transparente.

Seymour e colaboradores (2003) ressaltaram que a classificação foi apenas uma hipótese elaborada a partir de consensos na literatura de que determinadas línguas se organizam de forma diferente no que se refere à relação fonema-grafema e à estrutura silábica. Os autores apontaram a necessidade de uma análise linguística computacional de diversas línguas para classificar, de fato, as ortografias. Mesmo após 11 anos deste estudo, ainda não há, até o conhecimento da autora desta tese, um trabalho que tenha realizado esta análise linguística computacional. Futuros trabalhos poderão, com o auxílio de novas tecnologias, realizar esta análise que beneficiará a linguística e os estudos sobre ensino e aprendizagem da leitura e da escrita.

Soares (2018), por meio de uma reflexão sobre esta categorização de Seymour e colaboradores (2003), concluiu que a ortografia do português pode ser considerada de relativa transparência e está mais próxima à transparência. Em outras palavras, na ortografia do PB, há regularidades diretas (biunívocas) entre sons e grafemas, como os sons [p] e [b] na escrita de <p> e <b> em “pato” e “bola”; regularidades contextuais, como o som [k] na escrita de <c> diante de <a>, <o> e <u>, como “cama”, “cola”, “cuca”; e as irregularidades, ou seja, não há regras que regulam a relação de determinados fonemas e grafemas, como o fonema [s], que pode ser grafado pelo <s>, <ss>, <c>, <ç>, <xc>, como em “selo”, “assado”, “paçoca”, “exceto”. Nesta perspectiva, a ortografia não é apenas um sistema arbitrário, o qual precisa ser

totalmente memorizado, mas sim um sistema regulado por regras, que podem ser regulares, irregulares e/ou relacionadas a diversos conhecimentos. Estas são apenas suposições sobre a classificação da ortografia do PB, estudos empíricos precisam ser realizados futuramente para uma precisa classificação da profundidade ortográfica da nossa língua.

Mesmo se tratando de uma classificação hipotética da profundidade ortográfica por Seymour e colaboradores (2003), assim como as inferências realizadas por Soares (2018) ao afirmar que o PB é de relativa transparência ortográfica e de estrutura silábica simples, não podemos desconsiderar que as características da ortografia de uma língua impactam na aprendizagem da escrita e da leitura. Estudos evidenciam que o percurso de aprendizagem da escrita ortográfica se difere a depender da profundidade ortográfica da língua (Majorano *et al.*, 2021; Russak; Zaretsky, 2022). Caravolas e colegas (2013) mostram que, em idiomas com ortografias mais opacas, as habilidades de aprendizagem progridem mais lentamente para as crianças do que em idiomas com ortografias mais transparentes.

Além disso, a leitura em ortografias opacas *versus* ortografias transparentes segue diferentes trajetórias, porque são inerentemente apoiadas em processos diversos e requer o desenvolvimento de habilidades específicas da alfabetização na primeira infância (Majorano *et al.*, 2021). Russak e Zaretsky (2022) conduziram uma investigação sobre o aprendizado da escrita ortográfica em diferentes ortografias. As autoras argumentaram que a organização da ortografia pode fornecer uma janela para a maneira como o aprendiz compreende as relações ortográficas. Por exemplo, no português brasileiro, por ser uma ortografia de relativa transparência, os aprendizes tendem a usar a relação direta entre fonema e grafema com mais frequência. Ao contrário do inglês, por exemplo, que trata de uma ortografia opaca, a morfologia tende a contribuir no período de aprendizagem da escrita por parte da criança (Seymour *et al*., 2003).

Nesta seção, foi destacado que a Língua Portuguesa, por ser de base alfabética, a escrita é constituída a partir de relações entre fonemas e grafemas. Além disso, foi discutido que a ortografia é uma importante tecnologia na comunicação escrita e, portanto, ter domínio deste objeto é um auxílio para que o indivíduo saiba utilizar a escrita nos mais diversos contextos sociais. Ademais, as relações ortográficas foram apresentadas e vimos que, no português brasileiro, há casos regulares, regulares contextuais e irregulares. Por fim, discutimos que a profundidade ortográfica e a estrutura silábica de uma língua interferem diretamente no percurso de aprendizagem da escrita. Mesmo com a falta de uma pesquisa empírica, podemos inferir que a profundidade ortográfica do português brasileiro é de relativa transparência.

A organização das relações ortográficas – regulares, regulares contextuais e morfológicas e irregulares – também está presente nos documentos educacionais brasileiros, como a Base Nacional Comum Curricular – BNCC (Brasil, 2017, 2018).

3.2 Ortografia na Base Nacional Comum Curricular (BNCC)

O objetivo desta seção é discutir sobre o lugar da ortografia em um importante documento da educação brasileira em vigor, a Base Nacional Comum Curricular – BNCC (Brasil, 2017, 2018). Em um primeiro momento, apresentamos o que é a Base e como ela está organizada. Em seguida, identificamos os trechos em que a ortografia aparece neste documento.

A BNCC (Brasil, 2017, 2018) tem o objetivo de definir o "conjunto orgânico e progressivo de **aprendizagens essenciais** que todos os alunos devem desenvolver ao longo das etapas e modalidades da Educação Básica" (Brasil, 2017, p.7, grifo nosso). A Base está organizada em dois documentos, de acordo com os segmentos educacionais. O primeiro documento, referente à Educação Infantil, Ensino Fundamental Anos Iniciais (Ensino Fundamental I) e Anos Finais (Ensino Fundamental II), foi publicado em 2017; o segundo documento foi publicado em 2018, referente ao Ensino Médio (EM).

No âmbito do componente da Língua Portuguesa, a BNCC do Ensino Fundamental (Brasil, 2017) sugere que a ortografia seja trabalhada a partir do 3º ano do EFI, na prática da "Análise Linguística/semiótica (Ortografização)". Nos anos anteriores, 1º e 2º anos, o foco é dado aos aspectos da alfabetização por meio da prática "Análise Linguística/semiótica (Alfabetização)". No quadro a seguir, há uma síntese de trechos nos quais aparecem conteúdos relacionados à ortografia.

Quadro 6 - Ortografia na BNCC do Ensino Fundamental

Ano	Prática	Objeto de conhecimento	Habilidades
1º	Análise linguística/semiótica (Alfabetização)	Construção do sistema alfabético e da ortografia	(EF01LP08) Relacionar elementos **sonoros** (sílabas, fonemas, partes de palavras) com sua representação escrita.
2º			(EF02LP04) Ler e escrever corretamente palavras com sílabas CV, V, CVC, CCV, identificando que existem vogais em todas as sílabas.
3º	Análise linguística/semiótica (Ortografização)		(EF03LP01) Ler e escrever palavras com correspondências **regulares contextuais** entre grafemas e fonemas – c/qu; g/gu; r/rr; s/ss; o (e não u) e e (e não i) em sílaba átona em final de palavra – e com marca

Ano	Prática	Objeto de conhecimento	Habilidades
4º			(EF04LP01) Grafar palavras utilizando regras de correspondência fonema-grafema **regulares diretas** e **contextuais**.
5º			(EF05LP01) Grafar palavras utilizando regras de correspondência fonema -grafema **regulares**, **contextuais** e **morfológicas** e palavras de uso frequente com correspondências **irregulares**.
6°	Análise linguística/semiótica	Fono-ortografia	(EF67LP32) Escrever palavras com **correção ortográfica**, obedecendo as convenções da língua escrita.
7º			
8º			(EF08LP04) Utilizar, ao produzir texto, conhecimentos linguísticos e gramaticais: **ortografia**, regências e concordâncias nominal e verbal, modos e tempos verbais, pontuação etc.
9º			(EF09LP04) Escrever textos corretamente, de acordo com a **norma-padrão**, com estruturas sintáticas complexas no nível da oração e do período.

Fonte: BNCC (Brasil, 2017, grifo nosso).

Ao observar o Quadro 6, percebe-se uma gradação no ensino da ortografia do 3º ao 5º ano. No 3º ano, espera-se que o aluno tenha conhecimento e consiga ler e escrever palavras com correspondências regulares. No 4º ano, espera-se que ele já escreva palavras com regularidades diretas e contextuais. Já no 5º ano, último ano do Ensino Fundamental I, o aluno precisa dominar as relações ortográficas regulares, regulares contextuais e morfológicas e as palavras irregulares de uso frequente (Brasil, 2017). Ou seja, a BNCC (Brasil, 2017) espera que, no 3º ano, as relações ortográficas sejam introduzidas e, nos anos seguintes, aprofundadas progressivamente.

A partir do 6º ano, primeiro ano do Ensino Fundamental II, e do 7º ano, conforme a BNCC (Brasil, 2017), espera-se que o aluno escreva palavras com correção ortográfica. No 8º ano, a ortografia deve ser utilizada, atrelada a conhecimentos gramaticais, para produzir textos. Já no 9º ano, último ano do EFII, o termo ortografia não é citado na habilidade; há apenas a menção de que o aluno precisa escrever de acordo com a norma padrão. Pode-se inferir, portanto, que, ao final do EFII, de acordo com a BNCC (Brasil, 2017), o discente escreva ortograficamente e domine as relações ortográficas.

Na BNCC do Ensino Médio (Brasil, 2018), as habilidades de Língua Portuguesa estão situadas nas competências específicas de "Linguagens e suas Tecnologias" (Brasil, 2018, p.468). Em relação ao conteúdo ortográfico, a BNCC do Ensino Médio (Brasil, 2018) apenas menciona o termo "ortografia padrão", como pode ser visto no Quadro a seguir.

Quadro 7 – Ortografia na BNCC do Ensino Médio

Prática	Habilidade
Leitura, escuta, produção de textos (orais, escritos, multissemióticos) e análise linguística/semiótica	(EM13LP13) Planejar, produzir, revisar, editar, reescrever e avaliar textos escritos e multissemióticos, considerando sua adequação às condições de produção do texto, no que diz respeito ao lugar social a ser assumido e à imagem que se pretende passar a respeito de si mesmo, ao leitor pretendido, ao veículo e mídia em que o texto ou produção cultural vai circular, ao contexto imediato e sócio-histórico mais geral, ao gênero textual em questão e suas regularidades, à variedade linguística apropriada a esse contexto e ao uso do conhecimento dos aspectos notacionais (**ortografia padrão**, pontuação adequada, mecanismos de concordância nominal e verbal, regência verbal etc.), sempre que o contexto o exigir.

Fonte: BNCC (Brasil, 2018, grifo nosso).

Como pode-se observar no quadro anterior, de acordo com a BNCC (Brasil, 2018), espera-se que o aluno do Ensino Médio tenha domínio da ortografia e seja capaz de escrever ortograficamente em diversos contextos sociais. Podemos concluir que a ortografia está, de fato, presente na BNCC. Em relação aos Parâmetros Nacionais Educacionais – PCNs (Brasil, 1997), antigos parâmetros que regiam a educação brasileira, os quais criticaram o ensino mecânico da ortografia e propunha um ensino reflexivo da ortografia, a BNCC avança ao propor uma visão do ensino da ortografia no decorrer dos anos escolares.

No entanto, a BNCC, assim como os PCNs, não apresenta nenhuma orientação sobre como este conteúdo da ortografia pode ser trabalhado, além de não explicitar quais são as informações necessárias que o professor precisa saber para orientar o ensino da escrita ortográfica. Além disso, como veremos na seção seguinte, há uma discrepância entre a realidade da escrita ortográfica dos alunos em diversos níveis educacionais e o que é esperado pelos documentos da educação brasileira, como a BNCC (Brasil, 2018).

3.3 A escrita ortográfica de alunos brasileiros: da Educação Básica ao Ensino Superior

O objetivo desta seção é realizar uma revisão bibliográfica de trabalhos que investigaram a escrita ortográfica de alunos em diversos segmentos educacionais. Em um primeiro momento, discutiremos trabalhos desenvolvidos na Educação Básica brasileira, nos segmentos do Ensino Fundamental I – EFI – (1º ano ao 5º ano), do Ensino Fundamental II – EFI– (6º ano ao 9º ano) e do Ensino Médio – EM – (1ª à 3ª Série), e após esta, como no Ensino Superior. No segundo momento, afunilaremos nossa discussão sobre as irregularidades ortográficas, conteúdo no qual os indivíduos mais demonstram dificuldades na hora da escrita.

Como observamos na seção anterior, segundo a BNCC (Brasil, 2017), a partir do 3º ano do EFI, espera-se que o discente utilize conhecimentos linguísticos e gramaticais, como a ortografia, regras de concordância e pontuação, entre outros, na produção de textos. Portanto, infere-se que, a partir do 3º ano, as relações ortográficas são cobradas e ensinadas com mais ênfase na sala de aula. Além disso, a Base (Brasil, 2017) espera que os alunos do 5º ano finalizem o primeiro ciclo do EF com a capacidade de escrever palavras seguindo as correspondências regulares, contextuais e morfológicas, além de conseguir grafar palavras irregulares de uso frequente. No entanto, diversos trabalhos apontam que os alunos do EFI não cumprem com o que é esperado na BNCC (Brasil, 2017).

Santos e Befi-Lopes (2013), em uma pesquisa quantitativa, caracterizaram a ortografia de alunos do 4º ano de escolas públicas e particulares da Grande São Paulo. As autoras realizaram um ditado com os participantes, com palavras e não palavras, e contabilizaram os erros cometidos por estes alunos. Os resultados indicaram que, mesmo os alunos estando em um ano escolar pós-alfabetização, ainda foram encontrados altos índices de erros ortográficos cometidos pelo apoio à oralidade. Santos e Befi-Lopes (2013) argumentaram que apoio à oralidade é eficaz na alfabetização, mas os resultados indicaram que este apoio ainda é utilizado em um ano escolar que espera um certo domínio ortográfico dos alunos. Alguns exemplos de erros ortográficos que foram identificados são os seguintes: *infami (infame); “zurrar (surra); *combreti (complete); *resvilar (resfilar). As autoras concluíram que o erro ortográfico faz parte do processo de aprendizagem, mas que há a necessidade de um trabalho reflexivo com a ortografia.

Souza, Brandão e Melo (2020), em uma investigação qualitativa sobre a escrita de alunos do 5º ano do EF de uma escola pública, concluíram que a ortografia é uma das maiores dificuldades dos discentes no ato da produção textual. As autoras, por meio de um questionário destinado aos professores, investigaram quais seriam as maiores dificuldades dos alunos a partir da perspectiva dos docentes. Como resultados, Souza, Brandão e Melo (2020) identificaram, a partir da resposta dos professores, que os alunos possuem mais dificuldades em relação às regularidades contextuais e às irregularidades ortográficas. Além disso, elas relataram que há uma falta de trabalho reflexivo da ortografia na sala de aula. Entretanto, tais resultados podem ser questionados, pois foram apenas as percepções dos professores e não houve, de fato, uma análise dos textos dos alunos para afirmar que a ortografia é o maior problema de escrita encontrado nos textos dos discentes.

Diferente de Souza, Brandão e Melo (2020), que realizaram uma pesquisa qualitativa sobre a escrita dos alunos a partir das respostas dos professores, Zacharias-Carolino e Osti

(2020) investigaram, por meio de um ditado de diagnóstico, o desempenho ortográfico de alunos do 3º, 4º e 5º anos do EFI de uma escola pública. As autoras concluíram que os alunos têm um baixo desempenho em relação aos aspectos notacionais da escrita, pois apresentaram alto índice de erros na ortografia e na escrita alfabética, como *muindo (muito), *ezenpro (exemplo), *zelado (zelador), *carosa (carroça). A maior porcentagem de erro foi devida às representações múltiplas de grafemas, ou seja, as relações irregulares entre fonema e grafema, seguidas por erros relacionados à omissão de letras e à generalização de regras, entre outros. Os dados coletados pelas autoras indicam que, no decorrer dos anos escolares, o número de erros na escrita reduziu. No entanto, mesmo assim, Zacharias-Carolino e Osti (2020) ressaltaram a importância do trabalho com a ortografia, pois, mesmo com a redução de erros no passar dos anos escolares, os alunos ainda apresentam dificuldades nos anos finais do Ensino Fundamental, além daquelas esperadas pela BNCC.

No Ensino Fundamental II (EFII), Ferreira e Busse (2019) analisaram, qualitativamente, a ocorrência de fenômenos fonológicos em produções escritas dos alunos do 6º ano do EFII. As autoras coletaram produções textuais de discentes deste ano escolar de uma instituição pública e catalogaram os erros ortográficos. Ferreira e Busse (2019) encontraram erros, como a elevação ou abaixamento de vogais, como *isperta (esperta), *ermão (irmão) e a hipossegmentação, *quemora (que mora) e hiperssegmentação, *em bora (embora). Tais erros são esperados na fase de alfabetização, por ser um apoio à oralidade. No entanto, as autoras concluíram que eles persistem mesmo após a fase da alfabetização. Ferreira e Busse (2019) concluíram que os fenômenos fonológicos não se limitam apenas à fase da aquisição da escrita, pois tais erros perduram por toda a Educação Básica. A BNCC (Brasil, 2017), por exemplo, espera que tais erros tenham sido superados no 6º ano do EF.

Em uma pesquisa quali-quantitativa, Santos e Soares (2020), por sua vez, realizaram uma análise da escrita nos textos de alunos do 6º ano (primeiro ano do EFII) e 9º ano (último ano do EFII) de uma escola pública. Os discentes foram instruídos a elaborar duas produções textuais. Os erros encontrados nestes textos foram transcritos e catalogados. Os autores focaram em dois tipos de erros: (1) a interferência da fala na escrita, como monotongação e ditongação e alçamento vocálico, e (2) erros na convenção ortográfica. Os autores identificaram que os dois tipos de erros ortográficos tendem a diminuir com o passar dos anos escolares, mas que, mesmo ao final do Ensino Fundamental, a taxa de erros ortográficos ainda é expressiva.

A partir de uma pesquisa qualitativa, Nunes, Santos e Barbosa (2020) também analisaram os desvios de ortografia em produções textuais de alunos do 7º ano do EFII de uma escola pública. As autoras classificaram os desvios em dois grupos, a saber: (1) os decorrentes

da convenção ortográfica; (2) os desvios decorrentes da transposição de hábitos da fala para a escrita. Nunes, Santos e Barbosa (2020) encontraram frequentes erros ortográficos nos textos dos alunos, como a hipossegmentação, *quieu (que eu), o alçamento vocálico, por exemplo, *intão (então), troca de letras, *vique (fiquei), entre outros. Além disso, nos dados, foram encontrados desvios em palavras com irregularidades ortográficas, como *asustado (assustado), *comesei (comecei), *fexei (fechei), entre outros. A investigação das autoras demonstrou que os alunos cometem diversos tipos de erros ortográficos mesmo estando no 7º ano, ou seja, nos anos finais do Ensino Fundamental.

Também no EFII, Paula (2021) investigou os desvios ortográficos mais recorrentes em produções textuais de alunos do 9º ano. A autora constatou uma alta frequência de erros, como o apagamento do -R final em coda[4] de verbos no infinitivo, *esta (estar), na concorrência de grafemas para representar a fricativa /s/, *comesou (começou) e no alçamento vocálico, *discuti (discute). Além destes três tipos de erros, foram encontrados outros tipos, como erros na concorrência de grafemas para representar a fricativa palatal desvozeada /ʃ/, como *enxente (enchente) e *enfaicha (enfaixa). Paula (2021) concluiu, assim como Nunes, Santos e Barbosa (2020), que há uma necessidade de ações interventivas para tratar as dificuldades ortográficas que ainda permanecem nos textos de alunos, mesmo estando no ano final do EFII.

Em relação à ortografia no Ensino Médio (EM), Sartori, Mendes e Costa (2015), em uma pesquisa qualitativa, analisaram os erros ortográficos em produções textuais realizadas por alunos do EM de uma escola pública. As pesquisadoras identificaram que, mesmo ao final da Educação Básica, os alunos cometeram erros em vários tipos de relações ortográficas. As autoras identificaram erros em casos de regularidades diretas, como *tevo (devo) e *defemos (devemos); em regularidades contextuais, como *inprego (emprego) e *erado (errado); em regularidades morfossintáticas, como *ajuda (ajudar) e *escrevel (escreveu); e em irregularidades, como *geito (jeito) e *cituação (situação). Além disso, Sartori, Mendes e Costa (2015) afirmaram que os erros ortográficos em palavras irregulares são de maior ocorrência em detrimento aos erros em relações regulares.

Em uma pesquisa quantitativa em três momentos da Educação Básica, 6º e 9º anos do EFII e o 3º ano do Ensino Médio, Teis-Adamante e Busse (2022) investigaram a natureza e a frequência de erros gráficos em textos de alunos de duas escolas públicas. Também por meio de produções textuais de alunos de duas escolas, as autoras coletaram 271 textos e os erros ortográficos foram identificados e organizados em duas categorias, 1) Erros decorrentes das

[4] De acordo com Cristófaro Silva (2017, p.75), coda é o termo adotado para "indicar a parte pós-vocálica da sílaba".

assimetrias encontradas entre fonemas e grafemas; e 2) Erros com motivação marcadamente fonológica e/ou fonética. Em relação à categoria 1, os seguintes erros foram encontrados: *pasando (passando), *ouve (houve), *ansol (anzol), *foje (foge), entre outros. Na Categoria 2, os seguintes erros foram encontrados: *minina (menina), * mendingo (mendigo), * resouveu (resolveu), entre outros. As autoras identificaram que o número de erros ortográficos reduziu com o passar dos anos escolares. Além disso, Teis-Adamante e Busse (2022) identificaram que o erro ortográfico mais frequente se refere às relações ortográficas irregulares. Tais erros relacionados às irregularidades também reduziram ao longo do ano escolar. No entanto, mesmo ao final da Educação Básica, no 3º ano do EM, ainda houve uma porcentagem de erros relacionados às irregularidades ortográficas.

Também em três períodos distintos de escolaridade, Ensino Fundamental I e II e Ensino Superior, Nascimento e Henz (2020) analisaram, por meio de uma análise qualitativa, a escrita ortográfica em três produções textuais, cada uma de uma pessoa em diferente nível de escolarização: uma criança no 1º ano do EF I, um adulto que cursou o Ensino Fundamental II e um outro adulto do 6º período do curso de Letras. Na análise do texto da criança, os autores identificaram que ela está em um estágio avançado da alfabetização e, apesar de apresentar erros ortográficos, como troca de letras, como *apaijono (apaixonou); inversão e omissão de letras, *acordo (acordou), estes podem ser sanados por meio de intervenções. No texto do adulto alfabetizado, foram encontrados erros de troca de letras, como *elho (hélio); escrita fonética, como *disconficado (desconfiado); falta de acentuação gráfica e apagamento do -R em final de verbos. Os autores inferiram que este adulto, apesar de alfabetizado e de ter conseguido produzir um texto objetivo, demonstra um baixo conhecimento sobre a ortografia.

Na análise do texto do aluno da graduação, Nascimento e Henz (2020) identificaram erros ortográficos que não são esperados neste nível de escolarização, pois trata-se de um sujeito que finalizou a Educação Básica e está nos anos finais do Ensino Superior. Neste texto, os autores identificaram o apagamento do -R final em verbos, o uso de uma escrita fonética, como *mais (mas), *sutaque (sotaque); e o emprego inadequado dos grafemas concorrentes irregulares <c>, <ç>, <ss> e <x>, como *divercidade (diversidade), *compreenção (compreensão). Nascimento e Henz (2020) concluíram que os erros ortográficos apresentam semelhanças, como a escrita fonética, apagamento e omissão de letras, mesmo sendo em textos produzidos por pessoas em diferentes graus de escolarização e faixa etária. É importante ressaltar que o estudo desses autores foi realizado apenas a partir de um texto de cada um dos três indivíduos: uma criança e dois adultos, o que inviabiliza a generalização dos resultados. A

partir deste estudo, futuras pesquisas, com um maior número de participantes, poderiam investigar a escrita ortográfica de indivíduos em diferentes níveis de educação.

Em uma análise mista quali-quantitativa de erros ortográficos nas redações de candidatos do Exame Nacional do Ensino Médio (Enem) do ano de 2018, Oliveira, Couto e Lacerda (2023) encontraram erros de diversos tipos. Tais candidatos são concluintes do Ensino Médio, ou que já o concluíram, e estão em busca de uma vaga no Ensino Superior. As autoras identificaram erros que envolveram a troca de grafemas (*fazelidade para "facilidade"), o apagamento (*tira para "tirar"), a inserção de letra (*descidi para "decidi") e a inadequação do acento gráfico (*influênciando), entre outros. Oliveira, Couto e Lacerda (2023) concluíram que ainda há uma falta de domínio da escrita ortográfica mesmo tratando-se de pessoas que estão finalizando, ou finalizaram há anos, a Educação Básica. É importante ressaltar, também, que este estudo foi realizado a partir de um recorte de excertos de textos e redações completas que fazem parte do Manual do Corretor de Redação divulgado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Mesmo se tratando de apenas uma parte das redações do Enem, os tipos de erros ortográficos encontrados em textos de alunos concluintes, ou que já concluíram a Educação Básica brasileira, revelam falhas no processo de aprendizagem da escrita ortográfica.

Em relação ao Ensino Superior, Germani (2017) analisou, qualitativamente, a escrita de alunos do curso de Direito no último ano da graduação. Dentre os aspectos de escrita analisados pela autora, os erros de ortografia e de acentuação receberam destaque. Germani (2017) identificou uma realidade que destoa do esperado para o final do Ensino Superior, visto que foram encontrados diversos tipos de erros ortográficos nas produções escritas dos discentes. Foram encontrados erros de segmentação de palavras, como *em quanto ("enquanto"); apagamento do R-final em verbos, como *entra ("entrar"); e erros em palavras com irregularidades ortográficas, como *sensasionalista ("sensacionalista"), *salvacão ("salvação"), *infelismente ("infelizmente"), *sela ("cela"). Apesar de o foco de Germani (2017) não ter sido a escrita ortográfica, mas a produção de texto no geral, os erros ortográficos se destacaram dentre os demais erros de escrita. Futuras pesquisas poderão investigar a escrita ortográfica de alunos em diversas áreas do Ensino Superior.

A partir da discussão dos trabalhos acadêmicos desenvolvidos sobre a aprendizagem da ortografia de alunos brasileiros, podemos concluir que há, de fato, uma discrepância em relação ao que é proposto pela BNCC (Brasil, 2017, 2018) sobre o ensino da ortografia e a realidade da escrita ortográfica de alunos da Educação Básica brasileira. Os erros ortográficos analisados nos trabalhos relatados são, geralmente, semelhantes mesmo em diferentes períodos escolares.

Além disso, ao observar esses trabalhos sobre os erros ortográficos, dois aspectos semelhantes em toda a discussão podem ser destacados: o ano escolar e as irregularidades ortográficas. Há trabalhos que apontaram que o ano escolar contribui, de certa forma, para a redução dos erros ortográficos, mas, mesmo assim, ainda há uma defasagem em relação ao domínio da ortografia. Além disso, um dos pontos em comum destacados nos trabalhos discutidos é a dificuldade dos sujeitos participantes das pesquisas em relação às irregularidades ortográficas. Portanto, esta tese avança em relação à literatura precedente ao buscar analisar, qualitativa e estatisticamente, os anos escolares e as relações ortográficas irregulares com intuito de compreender a escrita de grafemas concorrentes em diferentes anos escolares, sendo 3º, 6º e 9º anos do Ensino Fundamental II e 3ª Série do Ensino Médio. Para assimilar melhor essas relações ortográficas irregulares, serão discutidos, na próxima seção, trabalhos acadêmicos que observaram os erros cometidos nas relações irregulares dos fonemas /s/, /z/, /ʃ/, /ʒ/ representados por grafemas concorrentes.

3.3.1 Irregularidades ortográficas: os fonemas /s/, /z/, /ʃ/, /ʒ/ e seus grafemas

O objetivo desta seção é caracterizar os fonemas /s/, /z/, /ʃ/, /ʒ/ e, por fim, discutir os trabalhos relacionados à aprendizagem dos grafemas que representam tais fonemas. A relação destes com os grafemas que os representam é classificada como irregular, pois não há nenhuma regra que regule qual grafema utilizar para representar determinado fonema. Por exemplo, na palavra “cidade”, não há uma regra para a escolha do grafema <c>, havendo uma concorrência entre o <c> e o <s>, pois ambos podem representar o fonema /s/.

Os fonemas /s/, /z/, /ʃ/, /ʒ/ são representados, acusticamente, pelos fones [s], [z], [ʃ], [ʒ]. Tais fones são classificados como fricativas, pois há uma fricção na passagem de ar pelo trato vocal (Cristófaro Silva, 2013), e sibilantes, pois apresentam “um sibilo durante a sua produção” (Cristófaro Silva, 2017). Os fones [s], [z], [ʃ], [ʒ] são, portanto, do ponto de vista articulatório, fricativas sibilantes. No Quadro 8 a seguir, apresentam-se as características articulatórias de cada uma das fricativas sibilantes do PB. Na primeira coluna, há os fones e, na segunda, as características articu

latórias destes.

Quadro 8 - Características articulatórias das fricativas sibilantes [s], [z], [ʃ], [ʒ].

Fones	Características articulatórias
[s]	Fricativa alveolar desvozeada

Fones	Características articulatórias
[z]	Fricativa alveolar vozeada
[ʃ]	Fricativa alveolopalatal desvozeada
[ʒ]	Fricativa alveolopalatal vozeada

Fonte: Cristófaro Silva (2013, p. 26-27).

Como observado no quadro anterior, os fones [s] e [z] possuem o mesmo modo de articulação, a fricção entre os articuladores, e o mesmo ponto de articulação, o alvéolo. Os dois sons se distinguem pelo vozeamento, a vibração das pregas vocais, sendo o [s] desvozeado e o [z] vozeado. O mesmo ocorre com os fones [ʃ] e [ʒ], ambos são fricativas e possuem o mesmo ponto de articulação, alveolopalatal, mas se distinguem pelo vozeamento, sendo o [ʃ] desvozeado e o [ʒ] vozeado.

Ortograficamente, os fonemas /s/, /z/, /ʃ/, /ʒ/ podem ser representados por diversos grafemas. No quadro a seguir, há uma descrição dos grafemas que podem representar essas fricativas. Na primeira coluna, estão os fonemas, na segunda, os grafemas que representam esses fonemas, por fim, na última coluna, os exemplos.

Quadro 9 – Os fonemas /s/, /z/, /ʃ/, /ʒ/ e seus grafemas.

Fonema	Grafema	Exemplo
/s/	<s> em início de palavra diante de <i> e <e>	**s**ino, **s**eda
	<c> em início de palavra diante de <i> e <e>	**c**ipó, **c**edo
	<ss>	a**ss**ento, fó**ss**il
	<c>	a**c**ento, dó**c**il
	<ç>	ru**ç**o, a**ç**úcar
	<sc>	cre**sc**er, cre**sc**imento
	<sç>	cre**sç**o, de**sç**o
	<x>	má**x**imo, sinta**x**e
	<xc>	e**xc**eção, e**xc**elente
/z/	<s>	ca**s**ar, me**s**a
	<z>	a**z**ar, **z**ebra
	<x>	e**x**ame, e**x**emplo
/ʃ/	<ch>	**ch**inelo, **ch**egar
	<x>	**x**ícara, li**x**o
/ʒ/	<j> diante de qualquer vogal	**j**iló, **j**ejum
	<g> diante de <e> ou <i>	**g**irafa, **g**elo

Fonte: Elaborado a partir de Soares (2018, p. 203).

Como observado no quadro anterior, os grafemas <s> e <c> concorrem para representar o fonema /s/ em contexto inicial de palavra diante de <e> e <i>. Isso quer dizer que no momento de aprendizagem da ortografia por parte do aluno, ele se depara com dois grafemas que podem

representar um mesmo som. Além disso, o contexto em que o fonema ocorre também não o auxilia na decisão do que escrever. Por exemplo, o fonema /s/, em contexto inicial de palavras, é representado pelos grafemas <s> e <c>, que são concorrentes tanto para quem está aprendendo a ortografia, quanto para as relações ortográficas. De modo semelhante, os fonemas /s/, /z/, /ʃ/, /ʒ/ estabelecem uma relação concorrente e irregular com os grafemas que os representam.

Por outro lado, o fonema /k/ estabelece uma relação regular contextual com os grafemas <c> e <qu>, ou seja, o contexto permite definir qual grafema poderá ser utilizado para representar o /k/. Por exemplo, o grafema <c> é utilizado diante de <a>, <o> e <u>, e o grafema <qu> diante de <e> e <i>. É importante ressaltar que, na perspectiva do aprendiz, <c> e <qu> também são concorrentes, pois ambos concorrem para representar o fonema /k/. No entanto, o contexto pode auxiliar o aluno na decisão de qual grafema escrever, se é o <c>, diante de <a>, <o> e <u>, ou o <qu> diante de <e> e <i>.

Tendo em vista que as irregularidades ortográficas são arbitrárias e não são reguladas por regras, há de se compreender que é uma parte difícil de aprender, pois não há nenhuma regra que motive a escrita de determinado grafema. Tal ideia pode ser vista em livros didáticos (LDs), que priorizam o ensino das relações ortográficas irregulares em detrimento das regulares (Teis-Adamante; Parise, 2018; Couto, 2020). Teis-Adamante e Parise (2018), por exemplo, observaram, em uma coleção didática de Língua Portuguesa amplamente utilizada nas escolas públicas brasileiras, que o enfoque dos LDs se refere à grafia de grafemas relacionados às fricativas [s], [z], [ʃ], [ʒ]. Ao encontro dos resultados de Teis-Adamante e Parise (2018), Couto (2020) constatou que os LDs do 6º e 7º anos do Ensino Fundamental privilegiam atividades que envolvem as exceções de palavras e as irregularidades ortográficas. Mesmo com o enfoque dos LDs nas relações ortográficas irregulares, estudos ainda apontam que uma das maiores dificuldades dos estudantes é em relação às irregularidades ortográficas (Sartori; Mendes; Costa, 2015; Nobile; Barrera, 2016; Marquardt; Busse, 2015; Saggiomo, 2018; Souza, 2019; Castro, 2022; Prado, 2023).

Em uma pesquisa qualitativa, Sartori, Mendes e Costa (2015) analisaram os erros ortográficos em produções textuais de alunos do Ensino Médio de uma escola pública. Os textos foram coletados e os dados catalogados. Os resultados da pesquisa identificaram que uma das maiores dificuldades na escrita ortográfica de alunos do Ensino Médio é com as relações irregulares entre fonema e grafema. As autoras encontraram erros ortográficos na relação entre os fonemas /s/, com os grafemas <c> e <s>, como *cituação (“situação”), e o / ʒ/, com os grafemas <g> e <j>, como *jeito (“jeito”).

Nobile e Barrera (2016), em uma pesquisa de abordagem quantitativa, investigaram as relações entre conhecimento ortográfico e produção escrita de textos de alunos do 5º ano de uma escola pública. As pesquisadoras coletaram os dados por meio de ditado e três produções escritas. Os resultados confirmaram, estatisticamente, que há uma correlação positiva entre escrever de acordo com a ortografia e produzir textos mais elaborados. Além disso, os resultados de Nobile e Barrera (2016) revelaram que os alunos, no geral, tiveram um conhecimento precário da ortografia, com frequentes erros ortográficos não apenas nas regularidades ortográficas, mas também nas irregularidades. As autoras trazem uma importante reflexão de que saber escrever de acordo com a ortografia pode, sim, estar correlacionado a escrever um bom texto. Portanto, estudos que investiguem as dificuldades dos alunos, como as relações irregulares ortográficas, são importantes para a compreensão destes empasses para a elaboração de práticas didáticas para o ensino reflexivo da ortografia.

Na mesma linha argumentativa de Nobile e Barrera (2016), Saggiomo (2018), em um estudo acerca de erros ortográficos de alunos do 3º e 6º ano do Ensino Fundamental, constatou que, nos dois anos escolares, a maior porcentagem de erros é destinada às irregularidades ortográficas. Saggiomo (2018) coletou os dados por meio de ditados realizados nos dois anos escolares. Os resultados indicaram, nos dois anos escolares, que a maior concentração de erros ortográficos ocorreu nas relações ortográficas irregulares, em casos como *sugeira (sujeira), * reflequiso (reflexo), *gosado (gozado), *insendio (incêndio) e, também, ao uso da acentuação gráfica.

Marquardt e Busse (2015), por sua vez, em um estudo sobre os erros ortográficos de alunos do 9º ano do EF, encontraram evidência de que a maior dificuldade, além dos erros ortográficos regulares, é nas relações irregulares ortográficas. As autoras afirmaram que a maior porcentagem de erros ortográficos depende da classe fonológica que o fonema pertence. Por exemplo, as autoras afirmaram que a classe das fricativas, como o fonema /s/, corresponde à maior dificuldade dos alunos, como, por exemplo, *centimentos (sentimentos).

Em uma pesquisa de abordagem qualitativa, Silva (2019) avaliou os efeitos do ensino de relações grafonológicas na redução dos casos de erros ortográficos em textos de alunos da modalidade Educação de Jovens e Adultos (EJA), no nível VI (8º e 9º anos do EFII). Um dos tipos de erros mais frequentes identificados pela autora refere-se à troca de grafemas concorrentes irregulares que representam o fonema /s/, como *ceguro (seguro) e *apareseu (apareceu).

Castro (2022), por sua vez, em uma pesquisa de abordagem mista quali-quantitativa, analisou os erros ortográficos mais persistentes na escrita e reescrita textual de alunos do 7º ano

do Ensino Fundamental e a relação entre o tipo de correção (resolutiva/indicativa) e a refacção do texto em relação à escrita ortográfica. Para isso, o pesquisador coletou uma primeira produção de textos dos alunos e realizou um tipo de correção, a resolutiva, em uma parte dos textos e a indicativa na outra parte. Nesta primeira produção textual, Castro (2022) identificou que a maior parte dos erros ortográficos, na tipologia troca de grafemas, se tratou de irregularidades ortográficas, especificamente de grafemas que representam fonemas /s/ e /z/. A seguir, há exemplos destes erros: *esperimentar (experimentar), *almoso (almoço), *feitisso (feitiço), *arros (arroz), *perceguir (perseguir), *pesquiza (pesquisa), *sosinho (sozinho), *vizualizações (visualizações), *cosinhar (cozinhar), entre outros. Após a correção, os alunos realizaram a reescrita e os erros ortográficos foram catalogados e analisados. Por meio de testes estatísticos, os resultados indicaram que não há relação entre o tipo de correção e a escrita de acordo com a ortografia. Além disso, os resultados apontam que, independentemente do tipo de correção, houve uma redução de erros ortográficos da primeira produção para a reescrita. Em relação à troca de grafemas, com relações fonema-grafema irregulares, também houve uma redução de erros ortográficos, mas, ainda houve erros nestas relações. Castro (2022) ressaltou a necessidade de um trabalho reflexivo e sistemático com a ortografia para que os erros ortográficos sejam superados.

Em textos de alunos do Ensino Superior, do curso de Direito, Germani (2017) identificou, por meio de uma pesquisa qualitativa, diversos erros ortográficos em palavras com irregularidades relacionadas às fricativas [s], [z], como *sensasionalista (“sensacionalista”), *salvacão (“salvação”), *infelismente (“infelizmente”), *sela (“cela”) e *dezeseis (“dezesseis”). Além disso, Nascimento e Henz (2021) identificaram erros ortográficos em produções textuais de alunos do Ensino Fundamental I e II e de um estudante do curso de Letras. Erros em várias palavras foram encontrados, como erros motivados pela oralidade e por processos fonológicos. Dentre os erros, foram identificados desvios em palavras com ortografia irregular, por exemplo, *ce (“se”); *apaijono (“apaixonou”) no texto do aluno do EFI; *nosa (“nossa”) no do EFII; e, por fim, *compreenção (“compreensão”), *excência (“essência”) e *divercidade (“diversidade”) no texto do aluno do curso de Letras. Mesmo em diferentes percursos escolares, Educação Básica e Ensino Superior, os discentes apresentam dificuldades na escrita de grafemas concorrentes irregulares.

Em uma pesquisa-ação de abordagem mista quali-quantitativa, Prado (2023) investigou o impacto de tipos de instrução para o ensino da ortografia em turmas do 6º ano do EFII de uma escola pública. A partir de uma primeira coleta de produções textuais e um ditado, a pesquisadora catalogou os erros ortográficos mais recorrentes e elaborou atividades didáticas

para o ensino da ortografia com três tipos de práticas, a saber: instrução direta; instrução contextual e espontaneísta. Após o desenvolvimento de cada tipo de prática em diferentes turmas do 6º ano, Prado (2023) realizou uma nova coleta de dados. Os resultados indicaram que os alunos se beneficiaram das práticas, pois houve uma redução de erros ortográficos no geral da primeira coleta para a segunda. No entanto, ao observar os erros ortográficos nas produções textuais em relação às irregularidades ortográficas, a pesquisadora observou que houve um aumento de erros em palavras de ortografia irregular. Os seguintes erros foram encontrados: *dacha, *decha ("deixa"), *fazendo ("fazendo"), *relaxante ("relaxante"), entre outros. Prado (2023), a partir deste resultado, ressaltou a necessidade de mais atividades pautadas em um ensino sistemático e reflexivo da ortografia.

Ao longo desta seção, as pesquisas discutidas mostraram que as relações ortográficas irregulares, principalmente dos fonemas /s/, /z/, /ʃ/, /ʒ/ com seus grafemas, configuram uma das maiores dificuldades na escrita de alunos ao longo da Educação Básica brasileira e do Ensino Superior. Podemos observar que a maioria dos trabalhos é de ordem qualitativa e são pautados na análise de textos ou de ditados, onde diversos tipos de erros ortográficos são discutidos. Esta tese avança em relação à literatura precedente ao propor, como objeto de pesquisa, o conhecimento ortográfico de alunos em diversos níveis da Educação Básica brasileira em relação à escrita de grafemas concorrentes irregulares. Além disso, esta pesquisa avança ao desenvolver um estudo experimental, com análises estatísticas dos dados, que investiga a escrita destes grafemas e a possível relação entre o ano escolar, a relação fonema-grafema e a frequência de ocorrência e de tipo podem influenciar o aluno no momento da escrita de palavras e não-palavras com irregularidades ortográficas. Dessa forma, esperamos que os resultados deste trabalho possam nortear a elaboração de atividades didáticas para o ensino das irregularidades ortográficas, além de contribuir para o avanço da compreensão de como funciona o processo, na perspectiva do aluno, da escrita de grafemas concorrentes irregulares.

3.4 Resumo do Capítulo 3

Neste Capítulo, foram discutidos os caminhos percorridos pela e na ortografia do português brasileiro. Na primeira seção, destaque foi dado à conceituação da ortografia e à sua importância para a comunicação escrita. Além disso, refletimos que a ortografia, como hoje conhecemos, passou por diversos períodos, como o fonético, o pseudoetimológico e, por fim, o período simplificado. A partir de um acordo entre Brasil e Portugal, unificou-se a ortografia e, em 2008, por meio do Decreto de Nº 6.583, foi promulgado, em terras brasileiras, o Acordo

Ortográfico da Língua Portuguesa. Por fim, as relações ortográficas foram apresentadas e foi possível observar que, na ortografia do PB, há casos de relações regulares, contextuais e irregulares entre fonema e grafema.

Nas segunda e terceira seções deste Capítulo, concluímos que há um descompasso entre o que a BNCC (Brasil, 2018) estipula para a aprendizagem da ortografia e a realidade da escrita ortográfica de alunos da Educação Básica e do Ensino Superior. Os trabalhos já desenvolvidos sobre a escrita demonstraram uma defasagem na ortografia de alunos em diversos anos escolares e, também, de discentes após a Educação Básica, já no Ensino Superior (Santos; Befi-Lopes, 2013; Souza, Brandão; Melo, 2020; Zacharias-Carolino e Osti (2020; Ferreira; Busse, 2019; Santos; Soares, 2020; Nunes; Santos; Barbosa, 2020; Paula, 2021; Sartori; Mendes; Costa, 2015; Teis-Adamante; Busse, 2022; Nascimento; Henz; 2020). Além disso, na terceira seção, foi destacada uma das principais dificuldades dos alunos, as relações irregulares entre fonema e grafema (Sartori; Mendes; Costa, 2015; Nobile; Barrera, 2016; Marquardt; Busse, 2015; Saggiomo, 2018; Souza, 2019; Castro, 2022; Prado, 2023). Neste Capítulo, foi possível observar a necessidade de investigar a escrita ortográfica de alunos em diferentes anos escolares em relação aos grafemas concorrentes irregulares. No próximo Capítulo, discutiremos sobre o papel da frequência na linguística e no ensino e aprendizagem da ortografia.

4 A FREQUÊNCIA EM EVIDÊNCIA

Nos Capítulos anteriores, refletimos sobre a organização da ortografia e sua aprendizagem. Além disso, vimos que uma das maiores dificuldades de alunos da Educação Básica e do Ensino Superior diz respeito às irregularidades ortográficas. Esta tese se propõe a investigar se o ano escolar, a relação fonema-grafema e a frequência de tipo e de ocorrência influenciam na escrita dessas irregularidades. Portanto, é importante aprofundar a discussão sobre a frequência. Neste contexto, os objetivos deste Capítulo são: discutir sobre o papel da frequência na Linguística, em específico na aprendizagem da ortografia, e apresentar os corpora utilizados para a conferência da frequência das palavras e dos padrões ortográficos investigados nesta tese.

Para dar continuidade às discussões deste capítulo, é essencial conceituar o termo frequência, incluindo seus desdobramentos: frequência de ocorrência e frequência de tipo. De acordo com Cristófaro Silva (2017, p. 122), frequência é a "medida relacionada com a ocorrência de elementos em um domínio ou *corpus*". A frequência de ocorrência, também conhecida por *token frequency*, é o "número total de vezes" em que uma palavra, ou outro elemento, ocorre em um determinado *corpus* (Cristófaro Silva, 2017, p.122). Por exemplo, a palavra "não", no *corpus* Linguateca (Costa; Santos; Cardoso, 2008), possui uma frequência de 467.0900, ou seja, ela repete 467.090 vezes no *corpus* mencionado. A frequência de tipo, também conhecida por *type frequency*, "é o número de elementos ou tipos que compartilham uma propriedade específica" em um *corpus* (Cristófaro Silva, 2017, p.112). Por exemplo, a sílaba [pa] ocorre 65.555 vezes no *corpus* Vocabulário Ortográfico Comum da Língua Portuguesa – VOC (Ferreira *et al*., 2017).

Na seção seguinte, serão discutidos trabalhos com foco na frequência desenvolvidos na área da Linguística. Na segunda seção, trabalhos que consideram a frequência na aprendizagem da ortografia serão discutidos. Por fim, os *corpora* para a conferência da frequência de tipo e de ocorrência, utilizados nesta tese, serão apresentados.

4.1 A atuação da frequência em pesquisas científicas

A frequência tem sido considerada como variável de impacto em diversos trabalhos desenvolvidos em áreas da linguística, da psicologia cognitiva, como na psicolinguística, nos estudos fonológicos, em estudos sobre a aquisição da linguagem, entre outros (Pinheiro, 1996; Huback, 2007; Guimarães; Cristófaro Silva; Gomes, 2020). De fato, nosso cérebro é capaz de

buscar padrões em qualquer ambiente em que esteja (Treiman; Kessler, 2014). Podemos inferir que a frequência é o número de vezes que determinado evento ocorre. Neste contexto, o objetivo deste Capítulo é discutir trabalhos desenvolvidos nas áreas da psicologia cognitiva e estudos fonológicos que consideraram a frequência como elemento de observação.

Na área da psicologia cognitiva, Pinheiro (1996) desenvolveu um estudo de identificação da frequência de palavras que ocorriam em livros didáticos de crianças da pré-escola (hoje a educação infantil) e as séries iniciais do primeiro grau (hoje o Ensino Fundamental I) na cidade de Belo Horizonte -MG. Diante da falta de um levantamento da contagem de frequência de ocorrência no PB, o objetivo central da pesquisadora foi mapear esta informação a partir de textos com os quais os alunos das referidas séries tinham contato.

Pinheiro (1996) coletou todos os textos completos de 12 livros infantis utilizados em escolas da 1ª à 4ª série do 1º grau (equivalente hoje ao 1º ao 4º ano do Ensino Fundamental I) na cidade de Belo Horizonte – MG. A pesquisadora coletou, no total, 263.909 palavras e um vocabulário de 18.272. Os dados foram distribuídos por série escolar e organizados em palavras de alta, média e baixa frequência de ocorrência. Para a definição destes três níveis de frequência, Pinheiro (1996) calculou a proporção de cada palavra pelo número total de palavras por série escolar. Por exemplo, o número de vezes em que a palavra "cidade" apareceu nos livros da 1ª série foi dividido pelo número total de palavras referente a esta série. Se o resultado da proporção fosse maior que 0,02, a palavra seria de alta frequência; se fosse até 0,00800, seria baixa frequência. As palavras de frequência relativa deveriam ser inferiores a 0, 0199 e superiores a 0,0080 (Pinheiro, 1996).

Este *corpus* desenvolvido por Pinheiro (1996) tem uma notável relevância, pois até hoje ele é utilizado como referência em estudos que possuem a frequência de ocorrência como variável de análise (Pollo, 2008; Santos; Befi-Lopes, 2013). É importante, no entanto, questionar se as palavras que eram frequentes em textos dos livros dos alunos na década de 90, ainda são frequentes hoje, após mais de 26 anos da elaboração e divulgação do corpus de Pinheiro (1996). Além de as mudanças corriqueiras desenvolvidas com o passar dos anos, os livros infantis adotados pelas escolas também mudaram. O Programa Nacional do Livro Didático (PNLD), por exemplo, seleciona, a cada quatros anos, novas coleções que serão adotadas nas escolas brasileiras. Diante disso, nesta tese, *corpora* atuais, como o Vocabulário Ortográfico Comum da Língua Portuguesa –VOC– (Ferreira *et al*., 2017), o Léxico do Português Brasileiro – LexPorBR (Estivalet, 2019) e o LexPorBR – Infantil (Estivalet *et. al*, 2023), e condizentes com a faixa etária dos participantes são utilizados para a conferência da frequência de tipo e de ocorrência.

Nos estudos fonológicos, pesquisas, ancoradas na Teoria de Exemplares[5], também consideram a frequência como variável de análise (Huback, 2007; Guimarães; Cristófaro Silva; Gomes, 2020). Huback (2007), por exemplo, analisou os efeitos da frequência, de tipo e de ocorrência, nas representações mentais dos falantes do português brasileiro. A pesquisadora investigou os grupos de plurais terminados em -ão ("escrivão" e "escrivães"), -l ("sal" e "sais") e em ditongo -u ("degrau" e "degraus"). Por meio do banco de dados do Projeto Avaliação Sonora do Português Atual (ASPA) e do Novo Dicionário Aurélio Eletrônico, a pesquisadora realizou a conferência das variações de cada uma das terminações -ão, -l e o ditongo -u e suas variações no plural. Além disso, Huback (2007) coletou dados de fala espontânea e não espontânea com os plurais de análise.

Os resultados da investigação de Huback (2007) indicaram que a interação entre frequência de ocorrência e frequência de tipo pode ser uma consequência do uso da língua. Por exemplo, palavras de alta frequência de uso, mesmo que o tipo do plural fosse infrequente, conseguem se manter. A partir dos resultados de Huback (2007), as seguintes generalizações podem ser realizadas:

- Alta frequência de ocorrência + baixa frequência de tipo = prevalência da frequência de ocorrência;
- Baixa frequência de ocorrência+ alta frequência de tipo = prevalência da frequência de tipo;
- Baixa frequência de ocorrência + baixa frequência de tipo = prevalência do que mais ocorre na classe.

Também com foco no plural de palavras, Guimarães, Cristófaro Silva e Gomes (2020) estudaram a aquisição oral do plural irregular do português brasileiro nos casos de substantivos terminados em -ão, como "mão" e "leão", e em ditongo oral decrescente, como "chapéu" e "lençol". As autoras, apoiadas na Teoria de Exemplares, coletaram produções orais de alunos de 3 a 12 anos de escolas de Belo Horizonte. As pesquisadoras definiram a frequência de ocorrência e de tipo a partir da seleção de 30 palavras no *corpus* ASPA (Cristófaro Silva *et al.*, 2005). A partir do *corpus*, foi definida a terminação de plural -ões como a de maior frequência em substantivos terminados em <ão>; e a terminação -is como de maior frequência em substantivos terminados com o ditongo. Os resultados do estudo indicaram a interação entre frequência de ocorrência e frequência de tipo, ou seja, se a frequência de tipo for maior e ocorrer

[5]Segundo Cristófaro Silva (2017, p. 210), a Teoria dos Exemplares, modelo representacional adotado pela Fonologia de Uso, "sugere que as representações linguísticas contêm aspectos redundantes e que os efeitos de frequência de tipo e frequência de ocorrência são crucias para a construção do conhecimento linguístico".

em palavras de baixa frequência de ocorrência, aquela passa a ser o padrão geral da produção. As conclusões de Guimarães, Cristófaro Silva e Gomes (2020) se assemelham às de Huback (2007) ao confirmarem a interação entre frequência de tipo e de ocorrência no desenvolvimento das representações mentais dos falantes. Diferente destes dois trabalhos que analisam a fala, esta tese analisa a escrita de grafemas concorrentes irregulares na escrita de alunos em diferentes segmentos escolares, com o objetivo de investigar interação da frequência de tipo e de ocorrência, o ano escolar e a relação fonema-grafema.

Nesta seção, observamos que a frequência é um elemento de interesse em pesquisas na área da psicologia e fonologia, por exemplo. Além disso, vimos que a frequência de tipo interage com a frequência de ocorrência para a construção das representações mentais dos falantes. Na seção seguinte, o destaque é dado à frequência no contexto da aprendizagem da escrita ortográfica.

4.2 A frequência no ensino e aprendizagem da ortografia

Na Base Nacional Comum Curricular (BNCC) do Ensino Fundamental (Brasil, 2017), há o uso do termo "palavras frequentes" para definir, dentro de uma habilidade, que o aluno do 5º ano do Ensino Fundamental (EF) deve dominar, sendo a escrita ortográfica de palavras irregulares de uso frequente. Além disso, segundo a teoria da Integração dos Múltiplos Padrões - IMP (Kessler; Treiman; 2014), o arcabouço teórico desta pesquisa, a frequência é um elemento essencial no processo de aprendizagem da escrita ortográfica. Portanto, o objetivo desta seção é apresentar trabalhos que observem o impacto da frequência no ensino e aprendizagem da escrita ortográfica.

Santos e Befi-Lopes (2013), em uma pesquisa quantitativa sobre os erros ortográficos de alunos do 4º ano do EFI, desenvolveram um ditado com dez palavras de alta frequência (PAF), dez de baixa frequência (PBF) e dez pseudopalavras (PP). Para a escolha das palavras de alta e baixa frequência, as autoras utilizaram a contagem de frequência realizada por Pinheiro (1996), discutida na seção anterior. Os resultados de Santos e Befi-Lopes (2013) indicaram que o número médio de erros ortográficos em palavras de baixa frequência foi maior do que nas de alta frequência. Tal pesquisa, portanto, demonstra que há uma correlação positiva entre frequência e redução de erros ortográficos. No entanto, é importante refletir se as palavras frequentes para crianças do ano de 1994, 1995, 1996 (anos de coleta e escrita do trabalho de Pinheiro (1996)) também seriam frequentes para os alunos 19 anos depois.

Ribeiro e Martins (2020), por sua vez, estudaram a aquisição da escrita de alunos do 3º ano do EFI em relação à multiplicidade de representação gráfica do fonema /s/. Sabe-se que o fonema /s/ pode ser representado ortograficamente por nove grafemas (<s>, <c>, <ss>, <ç>, <x>, <sc>, <xc>, <sç> e <z>). Os autores coletaram dados escritos destes alunos por meio de ditado de palavras, ditado de frase e produções de texto. Os dados foram catalogados quanto a erros e acertos de palavras com o fonema /s/. Ribeiro e Martins (2020) identificaram que o maior número de erros foi relacionado aos casos de <sç>, <xc> e <sc>, que são de baixa frequência no PB. Como conclusão, os autores ressaltaram que a frequência de tipo foi relevante, ou seja, os casos <sç>, <xc> e <sc>, por terem uma menor frequência de tipo no PB, são suscetíveis a serem escritos com erros. Além disso, os pesquisadores identificaram que a palavra "floresça", por ter uma menor frequência de ocorrência no *Corpus* Brasileiro[6], apresentou 24 casos de erros. Os achados de Ribeiro e Martins (2020) são importantes, pois ressaltam a importância da frequência de tipo e de ocorrência na escrita de alunos do EFI. No entanto, é válido refletir que o *corpus* utilizado para a conferência da frequência é composto por palavras retiradas de teses, dissertações, jornais, biografias, horóscopo e roteiros de cinema, não sendo condizente com a idade dos participantes da pesquisa de oito anos em média. Não há, neste *Corpus*, um recorte de palavras frequentes ao público infantil, mas a sua utilização se deve ao fato de não haver, para o PB, *Corpus* disponível com esses dados.

Não apenas no português a frequência atua na escrita ortográfica do aluno, em outras línguas, com na língua francesa (Pacton *et al.*, 2001), na espanhola (*Nigro et al.*, 2014), na alemã (Fay; Hein; Ghayoomi, 2015) e na italiana (Iaia *et al.*2002), por exemplo, também foi constatada essa influência. Na língua francesa, Pacton e colaboradores (2001) investigaram a aprendizagem implícita de crianças da 1ª a 5ª série em relação às regularidades ortográficas. Segundo os pesquisadores, no idioma francês, algumas consoantes são frequentemente duplicadas, como <m> e <l>, e outras são raramente ou nunca duplicadas, como o caso de <c> e <d>. Em um dos experimentos, as crianças tiveram que escolher pseudopalavras que mais se pareciam com palavras reais e dentre estas não-palavras estavam casos com consoantes simples e duplas. Como resultado, o estudo demonstrou que as crianças escolheram mais pseudopalavras que apresentaram casos de consoantes duplas de alta frequência, ou seja, foram sensíveis a padrões de alta frequência.

No italiano, Iaia e colaboradores (2022) investigaram os efeitos da frequência silábica e do comprimento de palavra na aquisição ortográfica de crianças italianas do 1º ano. Para isso,

[6] Link do *corpus*: http://corpusbrasileiro.pucsp.br/cb/Inicial.html

os pesquisadores coletaram os dados por meio do ditado de 60 palavras, classificadas em grupos de palavras curtas e longas, e de alta e baixa frequência da primeira sílaba. As palavras e a frequência foram selecionadas a partir de um dicionário com dados estatísticos para italiano escrito e lido por crianças do Ensino Fundamental elaborado por Marconi e colaboradores (1993). Além disso, para a definição da frequência silábica, os autores utilizaram os valores de Stella e Job (2000). Os resultados indicaram que as crianças têm um melhor desempenho na escrita de palavras curtas do que nas longas. Além disso, Iaia e colaboradores (2022) identificaram que as crianças também se mostraram sensíveis à frequência das sílabas, ou seja, palavras longas que eram iniciadas com sílabas de alta frequência eram escritas corretamente pelos alunos mais novos. Ao passo que, com o decorrer do processo de aprendizagem, o impacto da frequência da sílaba já não é importante para os alunos mais velhos. De acordo com Iaia e colaboradores (2022, p.7, tradução nossa[7]), essa redução do impacto da frequência da sílaba pode ter ocorrido à medida que os alunos se "tornam mais eficientes no mapeamento fino de fonema para grafema". Estes resultados ressaltam a importância da frequência de tipo no processo de aprendizagem da escrita ortográfica, principalmente na escrita de crianças menores.

Nigro e colaboradores (2014) investigaram se a capacidade de aprendizagem implícita de crianças do 3º ano falantes do espanhol tem relação com as habilidades de leitura e escrita. Um dos resultados da pesquisa demonstrou que a exposição dos alunos às palavras escritas desempenha um papel no processo de escrever palavras familiares. Isso aponta que, quanto maior a frequência de exposição dos alunos a determinadas palavras, maior será a capacidade de escrevê-las. Fay, Hein e Ghayoomi (2015), no alemão, analisaram a relação entre a frequência da palavra e o desempenho ortográfico de alunos em diversos níveis educacionais. Os pesquisadores, através de uma investigação empírica, argumentaram que o desempenho ortográfico de alunos falantes de alemão da 2ª à 8ª série em relação às palavras de baixa e alta frequência tem uma alta correlação entre a frequência e a escrita correta de palavras. Em outras palavras, os resultados indicaram que quanto maior a frequência da palavra, maior a possibilidade de o discente escrever esta palavra corretamente.

A partir dos trabalhos apresentados, pode-se observar que o aprendiz da escrita se beneficia dos elementos que ocorrem com frequência na língua. Em uma pesquisa sobre a

[7] No original: "become more efficient in fine phoneme-to-grapheme mapping, the facilitation induced by the syllable frequency is no longer detectable".

aprendizagem estatística[8] e a ortografia de alunos pré-fonológicos[9], Treiman e colaboradores (2019) apontam que crianças, na mais tenra idade, antes mesmo do processo de alfabetização, já conseguem identificar diferenças entre a escrita e outros desenhos. Além disso, o aprendiz, durante o processo de aprendizagem da escrita, mobiliza informações que foram coletadas ao longo de sua exposição à escrita, ou seja, quanto o maior contato com determinada forma de escrita, maior será o domínio. Isso envolve, portanto, uma aprendizagem estatística. Por exemplo, o trabalho de Cristófaro Silva e Oliveira (2013) mostra que os alunos resolvem rapidamente a grafia de <e> em contexto átono final de palavras. Isso ocorre porque, na fala no PB, o falante produz recorrentemente o som de [i] neste contexto, como em pent[i] e pend[i]. Tal som, na escrita, é, categoricamente, representado pelo grafema <e>. Então, um padrão pode ser o som ou um grafema, mas pode ser, também, o contexto, como a posição átona final. Todas essas informações são mobilizadas pelo aprendiz no momento da escolha de qual grafema escrever. Como a escrita de <e> na posição átona final é frequente no PB, este rapidamente é aprendido pelos alunos. Isso mostra a sensibilidade do usuário em relação aos padrões da língua.

Treiman, Decker e Kessler (2018), ao pesquisar sobre a sensibilidade de adultos em relação a diferenças grafotáticas no inglês, ou seja, ao analisar o contexto silábico da palavra, qual grafema o participante da pesquisa poderia escolher, concluíram que os adultos são sensíveis aos padrões silábicos, uma vez que os participantes selecionaram pseudopalavras que apresentavam sílabas comuns às palavras reais do inglês. Esses achados vão ao encontro de Treiman (2020), pois a autora argumenta que uma criança, durante a aprendizagem da escrita, é capaz de produzir palavras que estão de acordo com as propriedades fonotáticas da língua inglesa.

Como observado ao longo desta seção, a frequência, de tipo e de ocorrência, é um elemento essencial no processo de aprendizagem da escrita ortográfica no português brasileiro e em outras línguas, como italiana, alemã e espanhola. No entanto, é preciso tecer reflexões sobre este papel da frequência. Como observamos, cada trabalho segue uma definição do que é palavra frequente e infrequente na língua. Além disso, alguns trabalhos partem de *corpora* distantes da faixa etária dos participantes da pesquisa para a definição de palavra de maior ou menor frequência. Neste contexto, com o intuito de suprir essas duas lacunas, esta tese buscou

[8] O termo aprendizagem estatística refere-se à aprendizagem implícita proposta pela IMP (Kessler; Treiman, 2014). Este termo será melhor desenvolvido no Capítulo 4 – Arcabouço Teórico.

[9] Pré-fonológico refere-se à pessoa que ainda não passou pelo processo de alfabetização, ou caso esteja passado pela alfabetização, ainda não consegue ligar as unidades sonoras da língua às unidades gráficas, grafemas e letras.

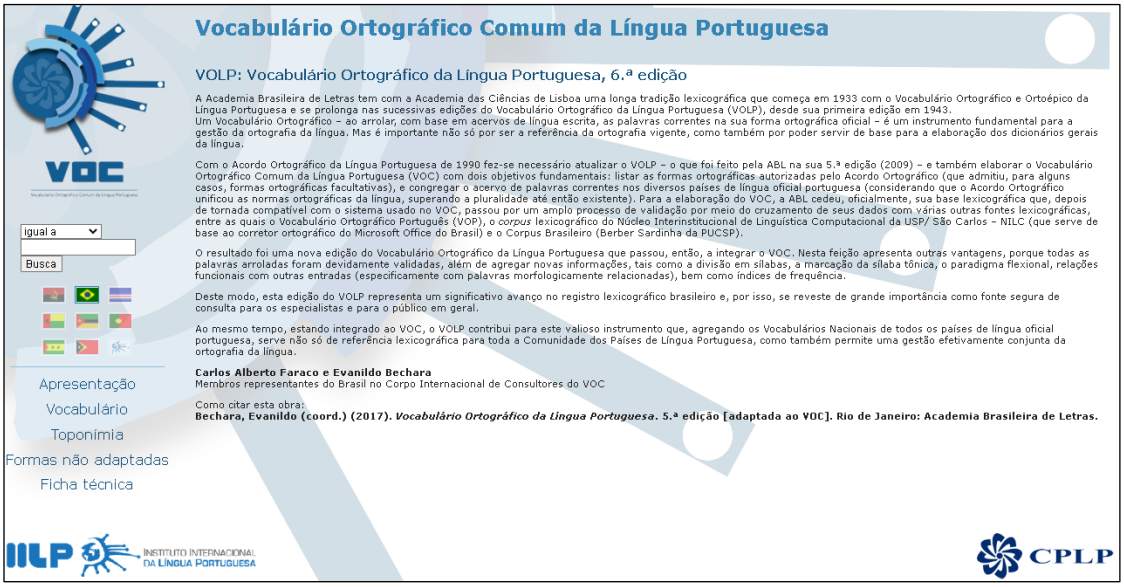
novos *corpora* que fossem atualizados frequentemente, além de corpus destinado a crianças, mais próximo à faixa etária dos participantes desta tese.

4.3 *Corpus* Vocabulário Ortográfico Comum da Língua Portuguesa (VOC)

A partir de busca sobre possíveis *corpora* que poderiam ser usados para a checagem da frequência de tipo de padrões ortográficos pesquisados nesta tese, o Vocabulário Ortográfico Comum da Língua Portuguesa –VOC– (Ferreira *et al*., 2017)[10] foi identificado. Portanto, o objetivo desta seção é apresentar o VOC, discutir suas características e limitações, além de realizar a checagem da frequência de tipo dos padrões ortográficos desta tese.

O VOC (Ferreira *et al*., 2017) é um importante instrumento para uma política de língua entre todos os países falantes do português. Em outras palavras, o VOC é uma base digital de dados que contém mais de dois milhões de formas ortográficas do português. Esta base está organizada de acordo com cada país, ou seja, é possível buscar as formas ortográficas utilizadas em todos os países falantes do português. No Brasil, o VOC (Bechara, 2017) é idealizado por Carlos Alberto Faraco e Evanildo Bechara, sendo este último o coordenador da base digital brasileira. A seguir, há a imagem da página inicial da plataforma referente ao Brasil.

Figura 1- Interface do Vocabulário Ortográfico Comum da Língua Portuguesa (VOC).



Fonte: Bechara (2017).

[10] Link de acesso ao VOC: https://voc.iilp.cplp.org/index.php

Ao lado esquerdo da página, de baixo para cima, há informações sobre a base de dados, como a apresentação do projeto idealizador da página, a ficha técnica, entre outras. Há também a opção para selecionar o país onde será realizada a busca da forma ortográfica. Por fim, há a opção de busca, que é constituída por três tipos, a saber: "igual a", "contém", "inicia com", "termina com". O usuário, por exemplo, pode buscar palavras que contenham <s>, por exemplo, que se iniciem com <s> e que se terminem com <s>.

O VOC, no entanto, precisa de melhorias, como a capacidade de diferenciar diacríticos, como o cedilha. Ele também não diferencia <s> e <ss>. Além disso, é impossível fazer a busca por meio do som, por exemplo, a busca por todas as palavras com o som [k]. Outra limitação da base é a impossibilidade em controlar algumas informações linguísticas, como o número de sílaba de palavras, a posição do acento e a diferença entre <c> e <ç>. Por exemplo, as palavras <canto> e <laço> são contabilizadas duas vezes como palavras que contêm <c>. No entanto, essas limitações não inviabilizam a importância do VOC (Ferreira *et al.*, 2017) para o acesso às formas ortográficas dos países lusófonos.

O VOC do Brasil (Bechara, 2017) foi escolhido como instrumento de coleta desta tese, pois, além de ser uma base com mais de dois milhões de formas ortográficas, ele está em constante atualização. Além disso, os recursos de controle de início e fim de palavra são essenciais para o escopo desta pesquisa. Por fim, a base permite saber a frequência de tipo de letras, sílabas, morfemas, entre outros. Por exemplo, 16911 palavras são iniciadas com o grafema <s> no português brasileiro. Além do VOC ser constantemente atualizado, não há, até o conhecimento da pesquisadora, nenhum outro *corpus* que seja possível fazer buscas e obter o número de ocorrências de padrões, sílabas e letras.

Nesta tese, destaque é dado a três relações irregulares de fonemas e grafemas, a saber:

- Fonema /s/, diante de <e> e <i> em início de palavra, pode ser representado pelos grafemas <c> ou <s>;
- Fonema /ʒ/, diante de <e> e <i>, pode ser representado pelos grafemas <g> e <j>;
- Fonema /z/, entre vogais, pode ser representado pelos grafemas <z> ou <s>.

A partir dessas relações nestes contextos, utilizamos o VOC (Bechara, 2017) para conferir a frequência de tipo das seguintes combinações: <ci>, <ce>, <si>, <se>; <gi>, <ge>, <ji>, <je>; <vzv>, <vsv>. Estes dois últimos são o <s> e <z> entre vogais. Na Tabela 1[11], a seguir, há o número de frequência dos grafemas <s>, <c>, <g> e <j> diante de <e> e <i> em contexto inicial de palavras e, por fim, a frequência dos grafemas <z> e <s> no contexto entre

[11] Um agradecimento à Tatiana Pollo pela rica discussão sobre a frequência de grafemas no português brasileiro.

vogais. Nas colunas da tabela, há os fonemas, o contexto de ocorrência, o grafema e sua frequência.

Tabela 1 – Frequência de tipo no VOC: fonemas e seus grafemas

Fonema	Contexto	Grafema	Frequência - grafema
/s/	Início de palavra diante de <e> e <i>	<c>	3682
		<s>	5666
/ʒ/	Diante de <e> e <i> em início de palavra	<g>	1476
		<j>	444
/z/	Entre vogais	<s>	110597
		<z>	92033

Fonte: Elaborado a partir do VOC (Bechara, 2017).

Como observado na Tabela 1, o fonema /s/, diante de <e> e <i> em início de palavra, ocorre 9384 vezes. Os grafemas <c> e <s> concorrem para representar o /s/, sendo o primeiro <c> frequência de 3682 e o <s> segundo como 5666. Como o VOC não permite a busca por som, pesquisamos cada uma das combinações dos grafemas <c> e <s> com o <i> e o <e>. Por exemplo, pesquisamos quantas vezes o <ci> ocorre em início de palavras, assim como o <ce>, e, também, o <si> e o <se>. Somamos, por exemplo, quantas vezes os padrões <ce> e <ci> ocorrem e chegamos ao resultado da frequência de tipo do grafema <c>.

O fonema /ʒ/, na Tabela 1, em contexto inicial de palavra diante de <e> e <i> ocorre 1920 vezes. Os grafemas <g> e <j> concorrem para representá-lo, sendo o primeiro com a frequência de tipo de 1476 e o segundo de 444. Reconhecemos que em outros contextos da palavra, também há a concorrência entre <g> e <j>, como em “sujeito” e “sargento”. No entanto, optamos por trabalhar apenas com o contexto inicial de palavras. No mesmo raciocínio do fonema /s/, pesquisamos quantas vezes o <gi> ocorre em início de palavras, assim como o <ge>, <ji> e <je>. Para obter a frequência de tipo dos grafemas <g> e <j>, em contexto específico, o número de vezes em que cada um dos padrões ocorre foi somado.

Além disso, podemos observar, na Tabela 1, que o fonema /z/ em contexto entre vogais pode ser representado tanto por <s> quanto por <z>. Realizamos as combinações dos grafemas <z> e <s> entre todas as vogais <a>, <e>, <i>, <o>, <u>. Por exemplo, pesquisamos quantas vezes as combinações <asa> e <aza> ocorrem. Ao todo, foram 50 combinações para chegar ao valor da frequência de tipo dos grafemas <s> e <z> em contexto intervocálico. Neste contexto,

o fonema /z/ pode ocorrer 202630 vezes, sendo a frequência de ocorrência de tipo de 110597 para o grafema <s> e 92033 para o <z>.

A frequência de tipo, um dos fatores analisados nesta tese, encontrada no VOC para cada um dos grafemas apresentados será utilizada para testar a hipótese de que a escrita ortográfica de grafemas concorrentes irregulares é influenciada por múltiplos fatores, como o ano escolar, a relação fonema-grafema, a frequência de tipo e de ocorrência. A partir da definição da frequência de tipo dos grafemas <c> e <s>, <g> e <j>, <z> e <s> nos contextos descritos anteriormente, as pseudopalavras[12] que compõem o experimento desta pesquisa, foram elaboradas. A descrição e explicação da elaboração destas não-palavras serão apresentadas no Capítulo 6 – Percursos metodológicos.

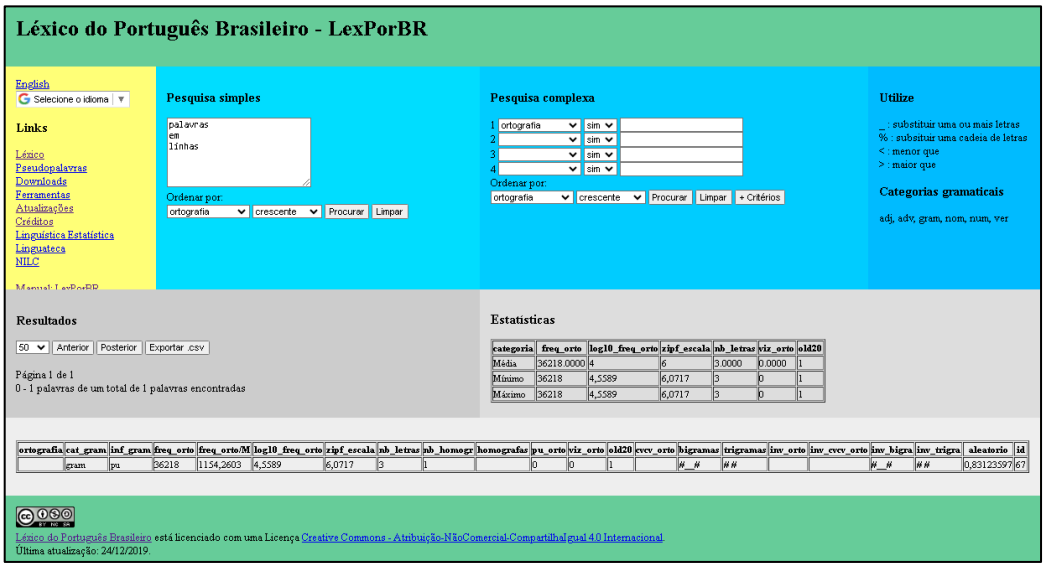
4.4 Corpora: Léxico do Português Brasileiro (LexPorBR) e LexPorBR - Infantil

Como observado nas seções anteriores, há impasses para a definição da frequência de ocorrência no desenvolvimento de pesquisas na Linguística. Com o intuito de reduzir esses impasses, novos *corpora* foram pesquisados e analisados: o Léxico do Português Brasileiro – LexPorBR[13] (Estivalet, 2019) e o LexPorBR – Infantil[14] (Estivalet *et. al*, 2023). Portanto, o objetivo desta seção é apresentar esses dois *corpora*, explicar as características de cada um deles e, por fim, elucidar a seleção de palavras utilizadas no experimento desta tese.

O LexPorBR é um *corpus* psicolinguístico livre e aberto que contém informações metalinguísticas e psicolinguísticas sobre palavras do português brasileiro – PB (Estevalet; Meunier, 2017). O *corpus* foi idealizado e desenvolvido por Gustavo Lopez Estivalet, em seu doutorado, diante da falta de *corpora* brasileiros que apresentassem informações importantes para a elaboração de pesquisas linguísticas. O LexPorBR é composto por mais de 32 milhões de palavras e possui mais de 215 mil entradas lexicais. A interface do *corpus* é de fácil navegação e organizada. A Figura, a seguir, mostra a página inicial do LexPorBR.

[12] Pseudopalavras "são sequências de caracteres que compõem um todo pronunciável, mas carente de significado" (Capovilla, Varanda, Capovilla, 2006, p.48). Em outros termos, pseudopalavras são não palavras inventadas para um determinado fim investigativo.
[13] Disponível no link: https://www.lexicodoportugues.com/
[14] Disponível no link: https://www.lexicodoportugues.com/infantil/

Figura 2- Página inicial do LexPorBR



Fonte: Estivalet (2019).

Na página do *Corpus*, há uma área para a pesquisa de palavras, além da formulação de pseudopalavras. Além disso, existe uma seção de Linguística Estatística com informações como a Distância de Levenshtein relativa, Vizinhos ortográficos, Distância de Hamming, entre outros. Ademais, há a opção de Downloads, que contém o manual e o *corpus* LexPorBR para baixar em diferentes versões, como modelo Txt., arquivo de Excel, entre outros.

Segundo Estevale e Meunier (2017), as palavras que compõem o LexPorBR foram retiradas de 13 arquivos *corpus* do Núcleo Interinstitucional de Lingüística Computacional (NILC) disponível no site Linguateca[15]. O NILC é um importante *corpus* do português brasileiro constituídos por outros *corpora*, elaborados a partir de diversos tipos de texto, como jurídico, didático, literário, técnico e cientifico, jornalístico e universitário (Pinheiro; Aluísio, 2003). As palavras retiradas dos arquivos do NILC foram organizadas em colunas de acordo com várias informações, tais como: frequência ortográfica (freq_orto), frequência ortográfica por milhão de palavras (freq_orto/M, [1000000*freq_orto/freq_total]), logaritm onatural da freq_orto/M (log10_freq_orto), número de letras(nb_letras), entre outras (Estevalet; Meunier, 2017).

Nesta busca por palavras nos LexPorBR, nos deparamos com o impasse de trabalhar apenas com um *corpus* que envolve palavras retiradas de textos diversos destinados, geralmente, a adultos. No entanto, os sujeitos desta pesquisa são de diferentes anos escolares e, consequentemente, diferentes faixas etárias, havendo alunos com 7 anos a 18 anos de idade.

[15] Link: https://www.linguateca.pt/acesso/corpus.php?corpus=SAOCARLOS

Diante desse contexto, foi realizada uma busca por um *corpus* que traria informações de palavras frequentes para as crianças. A partir disso, encontramos o LexPorBR – Infantil (Estivalet *et al*., 2023).

O LexPorBR – Infantil (Estivalet *et al*., 2019) foi desenvolvido com o objetivo de suprir a lacuna de *corpora* com foco no público infantil. Estivalet e colaboradores (2019) ressaltaram que a maioria dos *corpora* são elaborados a partir de textos escritos voltados para a população adulta, como jornais, revistas, entre outros. Nesse contexto, os autores desenvolveram o LexPorBR - Infantil. Este *corpus* é composto por um léxico de 130 milhões de *tokens* e 880 mil *types*. Ele é constituído por palavras retiradas de legendas de filmes e séries de comédia, família e animações em português brasileiro assistidos por crianças (Estivalet *et al*., 2019).

Ora, como utilizar um corpus elaborado a partir de palavras que as crianças apenas escutaram, sendo que a pesquisa foca na escrita ortográfica? Para responder a esta pergunta, é importante destacar que Estivalet e colaboradores (2019) verificaram que o vocabulário utilizado nas legendas também está de acordo com o encontrado em dicionários infantis sugeridos pelo Programa Nacional do Livro Didático (PNLD) para o 1º ano do Ensino Fundamental I ao 9º ano do Ensino Fundamental II. Portanto, mesmo sendo um corpus elaborado a partir de palavras que as crianças escutaram, textos escritos foram utilizados como basilares para a conferência da sua aderência ao vocabulário infantil. Isso torna o LexPorBR - Infantil relevante para esta pesquisa. Na Figura 3, a seguir, está a página inicial do *corpus*.

Figura 3 – Página inicial do LexPorBR – Infantil

Fonte: Estivalet e colaboradores (2023).

Diferente do LexPorBR, em que sua interface contém diversas informações e recursos, o LexPorBR – Infantil tem apenas o manual e os arquivos para *Download*, como dados, os

scripts no R. O arquivo com os dados está disponível em formato.Txt e pode ser lido pelo programa *Excel*. Os dados estão organizados em 48 categorias de informações, como lexema, lema, frequência da ortografia (Orto-freq), frequência da ortografia por milhão (Orto freq/M), frequência logarítmica (Orto freq log10), vizinhos ortográficos, entre outros.

Os recursos que constituem o *corpus* LexPorBR – Infantil (Estivalet *et al*., 2019), as informações de frequência de ocorrência de palavras e o fato dele ser direcionado ao público infantil, justificam o seu uso para a conferência da frequência de ocorrência e seleção de palavras para o experimento desta tese. Além disso, a diversidade de vocábulos do *corpus* o torna propício à conferência da frequência de ocorrência.

Dessa forma, com intuito de abranger todos os participantes desta pesquisa, da fase infantil à adulta, realizamos buscas de palavras nos dois *corpora*, o LexPorBR e o LexPorBR – Infantil. As palavras foram pesquisadas a partir dos fonemas e grafemas focos desta tese. Por exemplo, no padrão <ci>, duas palavras foram pesquisadas, uma de menor e outra de maior frequência de ocorrência nos dois *corpora* LexPorBR e o LexPorBR – Infantil. A palavra "cidade", por exemplo, tem frequência de ocorrência de 41918 no LexPorBR– Infantil e de 16.093 no LexPorBR. Ao passo que a palavra "cicuta" tem ocorrência de 63 vezes no *corpus* infantil e 18 no não infantil. Na tabela, a seguir, há as palavras selecionadas para o experimento desta tese. Na Tabela 2, há os fonemas, seguidos por seus grafemas, a palavra e a frequência de ocorrência nos dois *corpora*. Há quatro palavras com o grafema <c> e quatro com o <s> em relação ao fonema /s/. No par de palavras iniciado com <ci>, destacado em azul, pode-se observar que uma delas tem uma frequência de ocorrência menor em relação a outra nos dois *corpora*. Além disso, as duas palavras deste par possuem o mesmo número de sílaba (três sílabas), a mesma estrutura silábica (consoante + vogal), a mesma posição da tonicidade (paroxítona) e o grafema observado no início da palavra. Essas características, sempre que possível, foram controladas para a seleção de todas as outras palavras. Em resumo, as seguintes informações foram controladas para a seleção das palavras:

- Frequência de ocorrência similar em dois *copora*, LexPorBR e o LexPorBR – Infantil;
- Grafemas <c> e <s> diante de <e> e <i> em início de palavra;
- Grafemas <g> e <j> diante de <e> e <i>;
- Grafemas <s> e <z> entre vogais;
- Mesma posição na palavra, início, meio ou fim no par de palavras;
- Mesma posição quanto à tonicidade no par de palavras;

- Mesmo número de sílabas nos pares de palavras.

Tabela 2 – Palavras e a frequência no LexPrBR – Infantil e no LexPorBR – Não infantil

Fonema	Grafema	Palavra	Frequência de ocorrência	
			LexPor - Infantil	LexPor – Não infantil
/s/	<c>	cidade	41918	16093
		cicuta	63	18
		certo	312551	4668
		cerne	38	76
	<s>	semana	46151	14927
		sequela	49	21
		sistema	10879	10723
		singelo	26	40
/ʒ/	<g>	general	8051	1849
		gestual	19	33
		gigante	5847	206
		gincana	57	26
	<j>	sujeito	3667	1127
		dejeto	13	5
		jipe	552	114
		jiló	1	12
/z/	<z>	surpresa	17953	1220
		turquesa	123	14
		beleza	16301	1162
		leveza	107	115
	<s>	amizade	5813	608
		ojeriza	2	25
		camisa	8610	1207
		divisa	72	170

Fonte: Elaborada a partir de Estivalet (2019) e Estivalet e colaboradores (2023).

A frequência de ocorrência das palavras de ortografia irregular, um dos fatores analisados nesta tese, definida através do LexPorBR e do LexPorBR – Infantil será utilizada para o teste da hipótese de que a escrita ortográfica de grafemas concorrentes irregulares é influenciada por múltiplos fatores, como o ano escolar, a relação fonema-grafema, a frequência de tipo e de ocorrência. As palavras da Tabela 2, assim como o valor da frequência de ocorrência, farão parte do experimento desta pesquisa. A descrição do experimento será realizada no Capítulo 6 – Percursos metodológicos.

4.5 Resumo do Capítulo 4

Os objetivos gerais deste Capítulo foram discutir sobre o papel da frequência na Linguística, especificamente à aprendizagem da ortografia, e apresentar os *corpora* utilizados para a conferência da frequência das palavras e dos padrões ortográficos investigados nesta tese. Vimos que a frequência é assumida como variável de análise em trabalhos da psicologia e dos estudos fonológicos. Os estudos discutidos demonstraram que a frequência pode contribuir para a construção da representação mental dos falantes em relação aos sons da língua. Além disso, as pesquisas revelaram que pode haver uma integração entre a frequência de tipo e a frequência de ocorrência.

Em relação à aprendizagem da ortografia, os trabalhos encontrados na literatura concluíram que a frequência pode contribuir para um bom desempenho ortográfico de alunos de diferentes idiomas, como o português brasileiro, italiano, francês, alemão, inglês, entre outros. No entanto, foi discutido a necessidade de um olhar crítico para o que é considerado de baixa e alta frequência. Tal discussão teve como aporte a escassez de *corpora* do português brasileiro para a conferência de valores de frequência, e ao fato de os pesquisadores recorrem a corpus de dados antigos ou que não estão de acordo com a faixa etária dos participantes da pesquisa. Nesta circunstância, foram encontrados novos *corpora*, que são constantemente atualizados, e que estão em consonância à faixa etária dos sujeitos desta tese, o LexPorBR (Estivalet, 2019) e o LexPorBR – Infantil (Estivalet *et al.*, 2023) para a conferência da frequência de ocorrência.

A frequência, no entanto, não pode ser vista como solução para todos os problemas relacionados à aprendizagem da ortografia. No entanto, a partir das evidências dos trabalhos já realizados sobre a aprendizagem da escrita, a frequência, de tipo e de ocorrência, pode contribuir para a compreensão das dificuldades dos alunos em relação à escrita ortográfica. No próximo Capítulo, o arcabouço teórico desta tese será explicitado.

5 ARCABOUÇO TEÓRICO

O objetivo deste Capítulo é situar a tese no campo da Linguística Aplicada e apresentar a teoria guia desta pesquisa, a qual está inserida no campo da Linguística Aplicada, que abarca a "teorização em que teoria e prática sejam conjuntamente consideradas em uma formulação do conhecimento" (Lopes, 2008, p.101). Portanto, esta tese, no campo da Linguística Aplicada, assume uma concepção de língua em uso, pois é através do uso que são revelados e compreendidos conhecimentos do sistema linguístico (Treiman; Kessler, 2014), além de que "as circunstâncias de uso impactam a representação cognitiva da língua" (Bybee, 2016, p. 35).

O arcabouço teórico desta tese pauta-se em dois pontos essenciais, a saber: a revisão das teorias de aprendizagem da escrita ortográfica e a teoria que guia esta pesquisa. A seção "Teorias de aprendizagem da ortografia" refere-se à revisão das principais teorias relacionadas à aprendizagem da escrita e da ortografia. A segunda seção versa sobre a discussão da teoria da Integração de Múltiplos Padrões (IMP), postulada por Treiman e Kessler (2014), apontando como a IMP vai além das teorias discutidas na primeira seção.

5.1 Teorias sobre aprendizagem da escrita e ortografia

O objetivo desta seção é apresentar, discutir e compreender as teorias e os modelos teóricos de aprendizagem da escrita e da ortografia. A fim de maior clareza ao leitor, cabe a distinção entre escrita alfabética e escrita ortográfica. Nos anos iniciais da aprendizagem da escrita, o aluno se depara com o desafio de relacionar a fala à escrita, precisando aprender que: as unidades sonoras podem ser representadas por unidades na escrita; há espaço entre as palavras; cada letra tem suas características. Quando a criança se alfabetiza, ou seja, passa a ter conhecimento sobre o sistema de escrita, ela se depara com os desafios da ortografia (Ferreiro; Teberosky, 1991; Couto, 2020). Por exemplo, quando o aluno escreve "kaza" é sinal de que ele compreende o sistema alfabético de escrita do PB. No entanto, ainda precisa compreender as relações ortográficas para escrever "casa".

Na teoria da Memorização Mecânica, Jensen (1962) postula que o único caminho para a aprendizagem da escrita é por meio da memorização mecânica de palavras. Nesta perspectiva, esta memorização precisa ocorrer, pois não há regularidades na relação entre fonema e grafema que possam auxiliar o aprendiz na aprendizagem da escrita. Nesta teoria, pode-se inferir que o ensino da escrita é pautado na concepção de língua como estrutura, pois o foco é dado a atividades exaustivas, como cópia de palavras, preenchimento de lacunas (Castro; Couto,

2021). Ora, o aprendiz pode se beneficiar de características gerais da língua e realiza generalizações no ato de escrever, como foi discutido no Capítulo 4 – "A frequência em evidência". Há, sim, casos da necessidade de memorização de palavras, que não são usadas com frequência, mas é inviável generalizar que a escrita de cada palavra precisa ser memorizada.

Diferente da Memorização Mecânica, a Teoria de Fases, postulada por Ehri (2005), considera que há regularidades entre fonemas e grafemas que podem auxiliar o aprendiz no processo de aprendizagem da escrita. Nesta perspectiva teórica, a criança precisa relacionar o som às unidades da escrita. Por esta capacidade de relacionar sons e letras, segundo Cardoso-Martins e Corrêa (2008), a Teoria de Fases fundamenta-se no paradigma fonológico. A Teoria de Fases traz avanços em relação à aprendizagem da leitura e da escrita, pois postula as fases de aprendizagem, além de apresentar que há regularidades entre sons e letras. No entanto, como aponta Castro e Couto (2021), esta compreensão da relação direta entre fonema e grafema no paradigma fonológico precisa ser complementada com informações morfológicas e etimológicas para uma aprendizagem da escrita, principalmente em uma ortografia de relativa transparência, como a do português brasileiro.

Diferente da Teoria de Fases (Ehri, 2005), o Construtivismo, postulado por Ferreiro e Teberosky ([1984] 1991) na obra "Psicogênese da Língua Escrita", foca nos processos cognitivos das crianças no percurso da aprendizagem da escrita. No Construtivismo, as crianças criam hipóteses sobre a escrita a partir do contato com este sistema (Castro; Couto, 2021). Ferreiro e Teberosky ([1984] 1991) organizaram o aprendizado da escrita em cinco níveis, sendo o último, o estágio da alfabetização consolidada. As autoras salientam que é a partir do Nível 5 que a criança se deparará com dificuldades do sistema ortográfico. Mesmo não tendo o foco na aprendizagem da ortografia, a partir do Construtivismo, a consciência sobre o erro ortográfico como hipótese, e não como o peso negativo de errar, foi difundida. Entretanto, os níveis de aprendizagem da escrita propostos pelo construtivismo, assim como pela Teoria de Fases de Ehri (2005), podem ser questionados, pois a aprendizagem da escrita é múltipla e a criança pode avançar ou regredir nestes estágios (Pollo; Treiman; Kessler, 2015; Castro; Couto, 2021).

Outra importante teoria sobre a aprendizagem da leitura e da escrita é o modelo teórico da Dupla Rota (Coltheart *et al.*, 2001; Coltheart, 2006). Segundo Soares (2018), a Dupla Rota é amplamente difundida nos estudos sobre a leitura, mas não na escrita. Neste modelo teórico, a aprendizagem da leitura e da escrita ocorre por meio de duas rotas independentes, a rota sublexical (ou fonológica) e lexical (ou visual, ortográfica). Na primeira rota, o indivíduo acessa a relação entre fonema e grafema para conseguir realizar a leitura e a escrita. Caso esta rota não

seja acionada, entra em jogo a rota lexical, que envolve a memorização da palavra. No entanto, um erro ortográfico do aluno pode revelar que há diversas associações que ele pode realizar além de acionar apenas a rota fonológica ou a lexical. Por exemplo, no português brasileiro, se a aluno comete os seguintes erros *perata e *pepoca, ele realizou associações fonológicas e ortográficas. No PB, Toledo (2023) ressalta que palavras pronunciadas com [i] são escritas, frequentemente, com <e>, por exemplo, [miˈninʊ] -> "menina". Ao cometer o erro na escrita de "pirata" e "pipoca", o aprendiz pode ter realizado a associação de que ao pronunciar [i], a escrita precisa ser com <e>. Então, além da relação sonora, ele também pode ter seguido uma relação ortográfica. Portanto, é viável questionar se as rotas sublexical (ou fonológica) e lexical (ou visual, ortográfica) são independentes.

Um ponto em comum em todas as teorias e modelos teóricos discutidos até aqui é a separação de aprendizagem da escrita alfabética da ortografia. Em um primeiro momento, a criança aprende a escrita alfabética, por exemplo, para posteriormente aprender o sistema ortográfico. Com o avanço dos estudos linguísticos e de pesquisas sobre a aprendizagem da escrita, principalmente no âmbito da psicologia, ficou evidente que a criança beneficia da escrita, seja alfabética, seja ortográfica, desde o mais precoce contato com o mundo da escrita (Pollo, 2008; Treiman *et al.*, 2018; Otake; Treiman; Yin, 2017).

Por exemplo, Pollo (2008), em um estudo experimental, testou três diferentes visões sobre a aprendizagem da escrita, a abordagem fonológica, a construtivista e a estatística. Para isso, a pesquisadora analisou a escrita de alunos não fonológicos (pré-alfabetização) e fonológicos (em processo de alfabetização ou alfabetizados) falantes de português e inglês. Ao contrário do que a perspectiva fonológica postula, de que as primeiras escritas das crianças são letras arbitrárias, e do que o construtivismo postulou de que as primeiras escritas são de caráter universal e não refletem a língua com a qual a criança tem contato, Pollo (2008) demonstrou, a partir dos resultados de sua pesquisa, que as crianças, mesmo as pré-fonológicas, se beneficiam do conhecimento que adquirem ao serem expostas à escrita de sua língua. Em outras palavras, as crianças, desde a tenra idade, realizam a aprendizagem estatística em busca de padrões sobre a escrita da língua com a qual estão em contato.

Esta aprendizagem estatística é pautada na concepção de que as crianças, desde a tenra idade, conseguem captar, implicitamente, padrões em vários domínios (Treiman *et al.*, 2018). Otake, Treiman e Yin (2017), por exemplo, confirmaram que as tentativas de escrita de crianças chinesas e norte-americanas de dois a cinco anos de idade revelam características do sistema de escrita do indivíduo. Ao avaliarem essas tentativas, os adultos, falantes de chinês e inglês,

conseguiram julgar, com um significativo desempenho, a nacionalidade da criança que escreveu.

Neste contexto da aprendizagem estatística, em que o processo de aprendizagem da escrita alfabética e, também, ortográfica, é motivado por múltiplos padrões, surge a Teoria da Integração dos Múltiplos Padrões (doravante IMP), proposta por Treiman e Kessler (2014). Esta teoria avança em relação às teorias e modelos, discutidos anteriormente, ao propor uma visão múltipla do processo de aprendizado da escrita ortográfica. Na IMP, a memória é considerada constitutiva e, desde o início do contato com a escrita, a criança capta padrões sobre o sistema de escrita da língua em que está inserida.

Na IMP, utiliza-se o termo "phonograms", em português "fonograma", para se referir a unidades da escrita, como <f> e <b>. No entanto, nesta tese, o termo grafema é adotado para se referir a unidades mínimas e distintivas da escrita, como <c> e <b> em "casa" e "bata", <g> e <j> em "gato" e "jato". Esta escolha se justifica pelo termo grafema ser mais conhecido na área da Linguística Aplicada.

No Quadro 10, a seguir, há uma síntese das teorias ou modelos teóricos discutidos anteriormente. Na primeira coluna, está o nome da teoria/modelo teórico, seguido pelo seu idealizador e, após isso, uma breve descrição.

Quadro 10 - Síntese de teorias e modelos teóricos sobre a aprendizagem da escrita e ortografia

Teoria/Modelo teórico	Autor	Descrição
Memorização Mecânica	Jensen (1962)	A aprendizagem da ortografia ocorre por meio da memorização mecânica de palavras
Teoria de Fases	Ehri (2005)	Paradigma fonológico. Nesta perspectiva teórica, a criança precisa relacionar o som às unidades da escrita.
Construtivismo	Ferreiro e Teberosky ([1984] 1991)	A aprendizagem da leitura e da escrita é organizada em estágios.
Dupla Rota	Coltheart *et al.* (2001) e Coltheart (2006)	A aprendizagem da leitura e da escrita ocorre por meio de duas rotas independentes, a rota sublexical (ou fonológica) e a lexical (ou visual, ortográfica)
Integração dos Múltiplos Padrões (IMP)	Treiman e Kessler (2014)	O processo de aprendizado da escrita ortográfica é múltiplo e influenciado por múltiplos padrões

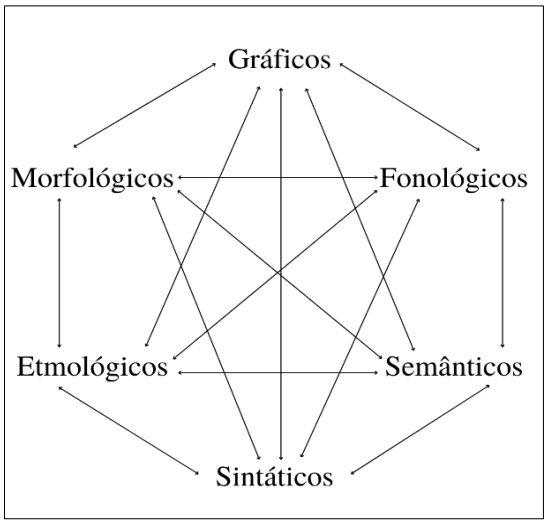
Fonte: Elaboração própria.

Na seção seguinte, a Teoria da Integração dos Múltiplos Padrões (IMP) será melhor apresentada e discutida.

5.2 Teoria da Integração dos Múltiplos Padrões (IMP)

Como discutido na seção anterior, a Teoria da Integração de Múltiplos Padrões (IMP) avança em relação à literatura precedente ao propor uma visão múltipla do processo de aprendizagem das relações ortográficas, além de reconhecer a ortografia como parte integrante deste processo. Na perspectiva da IMP (Treiman; Kessler, 2014), a ortografia, independente da língua, é motivada por múltiplos padrões. Segundo os autores, o indivíduo aprende a escrita por meio de duas categorias. A primeira refere-se ao padrão visual, em que ele consegue distinguir as letras e outros símbolos, assim como a noção de que estas possuem nomes e seguem um padrão fonotático. Enquanto a segunda categoria refere-se a padrões que estabelecem relação entre a forma gráfica e a linguagem, ou seja, os padrões ortográficos relacionados a conhecimentos linguísticos, como a fonologia, morfologia, sintaxe, semântica e etimologia. Dessa forma, a aprendizagem da escrita ortográfica envolve uma ampla rede de conhecimentos, que podem ser acessados pelo aluno no ato da escrita, como exemplificado na Figura 4.

Figura 4 - Rede de conhecimentos envolvendo a ortografia



Fonte: elaboração própria.

Em uma mesma palavra, podem atuar diferentes padrões. Por exemplo, na palavra “cidade”, há uma irregularidade na primeira sílaba, pois o som de /s/, diante de /i/, pode ser grafado como <s> ou <c>. Por outro lado, na última sílaba <de>, há uma regularidade, visto que o som de [i] em contexto átono final é sempre grafado com <e>. Além disso, a palavra “cidade” pode estabelecer relação com a palavra “cidadania”, “cidadão”, entre outras. O padrão morfológico pode ser mobilizado na escrita de se substantivos que são formativos a partir de

adjetivos. Por exemplo, os sufixos -eza e -ez formam substantivos de adjetivos, como “beleza”, de “belo”, e “maciez”, de “macio” (Soares, 2018).

A partir do exemplo dado anteriormente, pode-se observar que o conhecimento fonológico e o morfológico podem ser ativados para a escrita da palavra “cidade”, “beleza” e “maciez”. Neste contexto, a IMP compreende que a fonologia e a morfologia integram uma rede de conhecimento essencial para a apropriação da escrita, em que os padrões permitidos em determinada escrita estão ligados a estas duas áreas de conhecimento (Treiman *et. al.*, 2018).

Além de padrões fonológicos e morfológicos, como demonstramos anteriormente com a palavra “cidade”, os aprendizes também são sensíveis ao contexto, ou “posição condicionada[16]” (Treiman; Kessler, 2014, p.263, tradução nossa). Nesta perspectiva, a frequência de um grafema, por exemplo, pode modificar a depender dos segmentos vizinhos. Se a frequência muda, a sensibilidade do indivíduo em relação a ela também mudará (Treiman; Kessler, 2014). Além disso, como discutido no Capítulo 2 – “Panorama da pesquisa: objeto de estudo, hipótese e objetivos”, há trabalhos que confirmam que o contexto atua como importante informação no processo de aprendizagem do indivíduo (Toledo, 2023; Treiman; Kessler, 2014). Portanto, nesta pesquisa, dois contextos são considerados: os fonemas /s/ e /ʒ/ diante de <e> e <i> em início de palavra e o fonema /z/ entre vogais.

Para a IMP (Treiman; Kessler, 2014), há dois percursos de aprendizagem da ortografia, a implícita, guiada pela aprendizagem estatística, e a aprendizagem por instrução, guiada por instruções explícitas por outra pessoa. Segundo Pollo, Treiman, Kessler (2015, p.454), a aprendizagem estatística refere-se à frequência com que “um padrão ou uma regularidade estatística existe quando um grupo de eventos ocorre mais do que seria esperado pelo acaso”. A frequência está relacionada à “ocorrência de elementos em um domínio” (Cristófaro Silva, 2017, p.122). De acordo com Treiman (2018, p.644, tradução nossa), a aprendizagem estatística “permite-nos aprender sobre os padrões do mundo, ajudando-nos a prever eventos futuros e a nos comportarmos de forma adequada em novas situações”[17].

Trabalhos desenvolvidos na perspectiva da aprendizagem estatística ressaltam a importância da frequência da palavra e da frequência do padrão ortográfico. Treiman (1993), por exemplo, destacou que a criança falante de inglês é sensível ao tipo silábico e, desde cedo, já compreende que o uso de consoantes duplas <kk> ou <xx> não ocorre. As crianças são sensíveis, portanto, à frequência de grafemas individuais e à frequência dos pares de grafemas,

[16] No original: Conditioning by Position
[17] No original: It allows us to learn about patterns in the world, helping us to predict future events and to behave appropriately in new situations.

ou padrões ortográficos (Treiman; Kessler, 2022), ou seja, elas são sensíveis à frequência de tipo. Nesta tese, adotamos o conceito de frequência de tipo e frequência de ocorrência para analisar, atrelado ao ano escolar e a três casos de relação fonema-grafema, o processo múltiplo da escrita ortográfica de grafemas concorrentes regulares no português brasileiro. Isso trará novas reflexões sobre o papel da frequência na aprendizagem da ortografia.

Em relação à aprendizagem da leitura e da escrita ortográfica, a aprendizagem estatística é de suma importância para o reconhecimento de padrões que não foram ensinados (Treiman; Kessler, 2022). Portanto, na perspectiva da IMP, a frequência é importante, pois, quanto mais o aluno é exposto a determinados padrões ortográficos, mais robustas são as categorias de tais padrões. Há evidências de que até mesmo uma criança, que ainda não iniciou o processo formal de ensino e aprendizagem da escrita, consegue identificar os padrões (Treiman, 2017; Treiman *et al*. 2019). Por isso, o desenvolvimento da escrita deve ir além da memorização, ou seja, deve partir da relação entre grafia e sentido, de maneira que a frequência de uso e os padrões linguísticos fazem parte da construção do conhecimento (Treiman; Kessler, 2014).

Em uma análise de dados de escrita de crianças falantes de inglês, Treiman (1993) identificou que, mesmo ainda não estando no processo de aprendizagem da ortografia, as crianças demonstraram conhecimento sobre padrões da língua. Por exemplo, crianças no primeiro ano já não mais cometiam erros de consoantes duplas na mesma palavra, como <kk> ou <xx>, por não ser um padrão comum no inglês. Portanto, Treiman (1993) confirmou que as crianças nos estágios iniciais da aprendizagem da escrita conhecem padrões da língua. Em outras palavras, as crianças fazem uso da aprendizagem estatística desde a tenra idade.

Em um estudo sobre a aprendizagem estatística em crianças pré-fonológicas novas e mais velhas, Treiman e colaboradores (2018) identificaram que, antes mesmo de relacionar letras a sons, as crianças possuem conhecimento sobre a escrita e sobre os padrões da língua. Por meio destes dados, os autores demonstraram que a aprendizagem estatística ocorre antes mesmo do início da aprendizagem formal da escrita. Chetail (2017), por sua vez, investigou o impacto da sensibilidade do aluno a regularidades em situações de processamento visual de palavras. A autora identificou que, mesmo após um curto período, os indivíduos se mostraram sensíveis à frequência com que grupos de letras ocorrem mais do que outros. Nesse sentido, Treiman (2020) argumenta sobre a importância de pesquisas que quantifiquem padrões de sistemas de escritas e as estatísticas destes padrões a que os aprendizes são expostos. Além disso, Altmiller, Treiman e Kessler (2023) destacam que analisar o impacto do aprendizado estatístico ao longo dos anos escolares, considerando as variações na relação entre fonema e

grafema nas palavras, pode contribuir para a compreensão do desenvolvimento da aprendizagem ortográfica.

Apesar do positivo impacto da aprendizagem estatística na aprendizagem da escrita ortográfica, apenas ela não é suficiente para a compreensão da ortografia (Oliveira; Castro; Couto, 2023). Isso pode ocorrer, pois, essa aprendizagem "é lenta e depende muito do uso individual da língua" (Oliveira; Castro; Couto, 2023, p.49). Em outras palavras, metodologicamente, é difícil medir o quanto um indivíduo é exposto à escrita e qual a sua capacidade de realizar generalizações sobre a língua. Além disso, definir o que é frequente ou não na língua, e, consequentemente, à realidade do aluno, também pode configurar como um impasse para a realização de pesquisas na área. Por isso, esta tese buscou descrever padrões da ortografia do português a partir de *corpora* atualizados, como o VOC (Bechara, 2017), e condizentes com a idade dos sujeitos da pesquisa, como o LexPorBR (Estivalet, 2019) e o LexPorPR – Infantil (Estivalet *et al.*, 2023). Esperamos que, desta forma, a dificuldade de definir os níveis de frequência de palavras e padrões ortográficos e a experiência com a escrita de cada participante seja reduzida.

Além da aprendizagem estatística (implícita), é necessário, também, na perspectiva da IMP, as instruções explícitas por meio de outra pessoa sobre os padrões ortográficos. Assim, cabe ao professor expor para os alunos informações acerca da língua, orientá-los em relação à hipótese criada na aprendizagem implícita, para que eles possam reorganizar o seu conhecimento e aprender um novo padrão ortográfico (Treiman; Kessler, 2014).

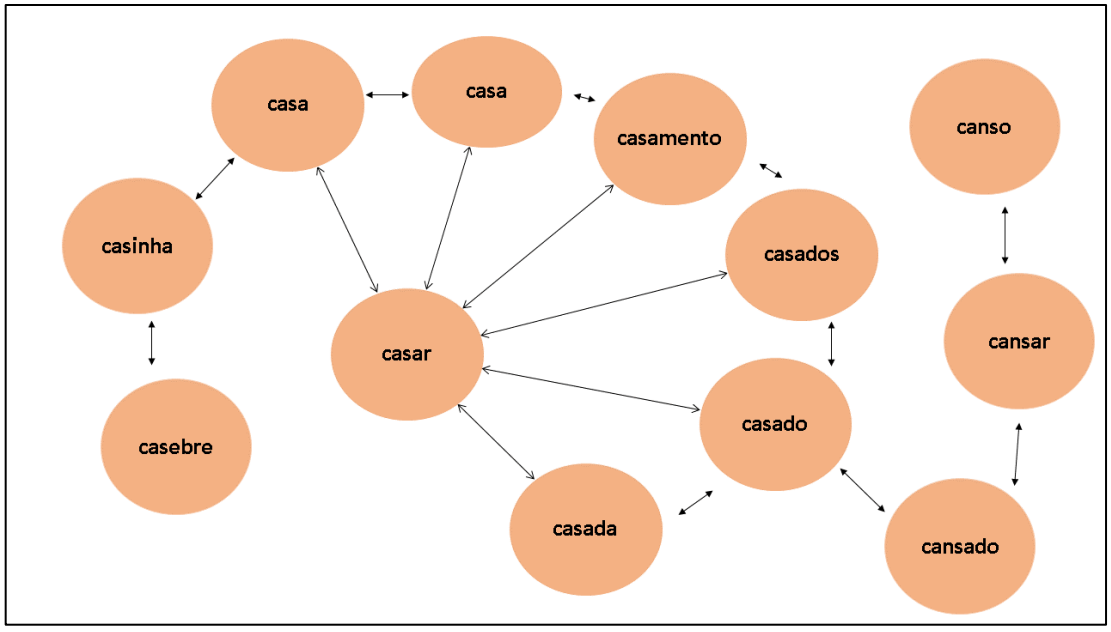
Graham e Santangelo (2014) realizaram uma meta-análise de trabalhos sobre instruções ortográficas na construção de bons leitores e bons usuários da escrita ortográfica. Os autores encontraram diversos trabalhos que confirmaram que o uso de instruções explícitas contribui para a formação do bom leitor e para o bom uso da escrita ortográfica. Prado (2023), por sua vez, demonstrou que os alunos do 6º ano do EFII de uma escola brasileira, com instruções diretas e reflexivas, melhoraram a escrita ortográfica em palavras isoladas e na elaboração de textos espontâneos. O trabalho de Prado (2023) evidencia, portanto, o benefício das instruções no processo de aprendizagem da ortografia. Dessa forma, os resultados de Graham e Santangelo (2014) e Prado (2023) vão ao encontro da IMP ao postularem a importância de instruções explícitas no processo de ensino e aprendizagem da escrita ortográfica.

Outro postulado importante na IMP é o fato de a memória ser vista como construtiva, e não apenas mecânica. Por meio da memória, o aprendiz da escrita ortográfica pode perceber e lembrar de eventos específicos, com generalizações que se desenvolveram com base na experiência com vários itens ou em generalizações que foram transmitidas por um adulto

(Treiman; Kessler, 2014). Para exemplificar estas generalizações, Oliveira, Couto e Lacerda (2023) identificaram erros de acentuação em textos de candidatos brasileiros do Enem 2018, como *mecânismo (“mecanismo”) e *econômias (“economias”). O sujeito, através de seu contato com as palavras “mecânico” e “econômica”, pode ter generalizado o uso do acento gráfico e, por isso, acentuou “mecanismo” e “economias”. Mesmo que o sujeito tenha transgredido uma norma de acentuação do PB, ele realizou generalizações e lembrou do acento gráfico.

Além disso, Castro e Couto (2021) apresentaram um exemplo de como trabalhar a memória construtiva a partir da construção de uma rede de palavras. Os autores, a partir da IMP, sugeriram como trabalhar com o erro ortográfico de troca de <z> por <s>, como em *cazar (casar). A partir de “casar” o professor pode desenvolver, com os alunos, uma rede de palavras com som de [z], mas escrita com <s>. Na Figura a seguir, há esta rede de palavras.

Figura 5 – Rede de palavras relacionadas ao vocábulo “casar”



Fonte: Castro e Couto (2021, p.1342).

Através desta rede de associações entre palavras apresentada na Figura 5, os alunos poderão compreender a escrita ortográfica irregular da palavra “casar”. O processo de aprendizagem da ortografia pode ocorrer de forma investigativa pela busca de palavras semelhantes e que compartilham padrões ortográficos similares (Treiman; Kessler, 2014). Cada padrão da palavra pode envolver diferentes tipos de unidades linguísticas, seja fonológica,

morfológica, gráfica, entre outros. Na Figura 5 a seguir, elaborada por Oliveira, Castro e Couto (2023), os múltiplos padrões podem ser observados em uma mesma palavra.

Figura 6 – Padrões da palavra desamarrotar

Fonte: Oliveira, Castro e Couto (2023, p.49).

Apesar de a IMP ser uma teoria desenvolvida na língua inglesa, ela é extensível à aprendizagem da ortografia de diversas línguas de base alfabética. No Brasil, a IMP se aplica e assume uma posição de extrema relevância ao trazer esta visão múltipla da aprendizagem da ortografia em um contexto em que esta assumiu uma posição de inferioridade nas aulas da Língua Portuguesa e, também, nos Estudos Linguísticos. Oliveira, Castro e Couto (2023, p.3) destacam que o ensino da ortografia passou "de um ensino extremamente mecânico e formal para um ensino voltado para o texto". O olhar para o texto, para o discurso, deixou de lado o ensino da ortografia e abriu espaço para o senso comum de que a ortografia não pode ser ensinada em palavras isoladas, ao passo que seu ensino deve ser apenas através do texto. Além disso, os pesquisadores que investigam a ortografia do PB e sua aprendizagem são, comumente, julgados pela ideia de que "estudar ortografia não é fazer linguística".

Neste contexto, a IMP assumiu o lugar de guia para as pesquisas sobre ortografia e seu ensino, principalmente no Grupo de Pesquisa sobre Práticas de Ensino de Escrita e Oralidade (PENSEO). Esse lugar foi assumido pois, na Teoria da Integração dos Múltiplos Padrões (Treiman; Kessler, 2014), a ortografia assume o seu lugar na linguística, pois sua aprendizagem ocorre na junção de formas gráficas a unidades linguísticas. Compreender que a ortografia é constituída por diversas relações, como fonológicas, morfológicas, etimológicas, fonotáticas, entre outras, identificar os processos por trás de um erro ortográfico, e como os estudantes manipulam as informações sobre a escrita e constroem generalizações sobre a ortografia a partir de sua experiência de uso da língua são caminhos para o fazer linguístico.

A IMP (Treiman; Kesller, 2014), portanto, subsidia esta pesquisa por considerar a singularidade do sujeito, além de compreender a aprendizagem da ortografia como um processo múltiplo e dinâmico em que o discente atua, diretamente, na construção de seu conhecimento. No quadro a seguir, há uma síntese dos principais pontos da IMP. Na primeira coluna do quadro, há os principais conceitos, seguidos por descrições.

Quadro 11 – Síntese da seção "Teoria da Integração de Múltiplos Padrões (IMP)"

Integração de Múltiplos Padrões	
Aprendizagem da ortografia	aprendizagem motivada por múltiplos padrões, que podem ser fonológicos, morfológicos, etimológicos, entre outros.
Memória	a memória é construtiva.
Aprendizagem estatística (implícita)	reconhecimento e uso de padrões.
Aprendizagem explícita	ensino de padrões por meio de uma outra pessoa.

Fonte: Elaborado a partir de Treiman e Kesller (2014).

5.3 Resumo do Capítulo 5

O Capítulo 5 situa esta tese no campo da Linguística Aplicada e introduz a teoria fundamental guia desta pesquisa. O objetivo foi situar esta tese no campo da Linguística Aplicada e apresentar a teoria guia desta pesquisa. Além disso, discutimos sore a concepção de língua em uso e propomos uma relação entre teoria e prática para entender esse conhecimento, especialmente no processo de escrita de grafemas concorrentes irregulares.

No decorrer do Capítulo, vimos que a aprendizagem da escrita foi investigada a partir de diversas teorias ou modelos teóricos, como a Teoria da Memorização Mecânica (Jensen, 1962), a Teoria de Fases (Ehri, 2005), o Construtivismo (Ferreiro; Teberosky, [1984] 1991) e a Teoria de Dupla Rota (Coltheart *et al.*, 2001; Coltheart, 2006). Em todas essas perspectivas, a ortografia não era considerada. Somente com a Teoria da Integração dos Múltiplos Padrões - IMP (Treiman; Kessler, 2014), a ortografia ganhou destaque no processo de aprendizagem da escrita desde a infância. Além disso, foi discutida a importância da frequência na aprendizagem da ortografia, assim como a importância da manipulação de diversos conhecimentos nessa aprendizagem. Segundo a IMP, a aprendizagem da ortografia é influenciada por diversos padrões, como fonotáticos, fonológicos, morfológicos, gráficos e, também, pela frequência com que o aprendiz é exposto à escrita da língua. O próximo Capítulo delineará o percurso metodológico adotado nesta pesquisa.

6 PERCURSOS METODOLÓGICOS

O arcabouço metodológico desta tese é pautado na abordagem mista quali-quantitativa (Paiva, 2019), do tipo de pesquisa de campo experimental (Lakatos; Marconi, 2003) e descritiva (Andrade, 2010). A abordagem mista quali-quantitativa permite a análise de aspectos subjetivos da pesquisa, combinada com o uso de testes estatísticos dos dados coletados. Além disso, o tipo de pesquisa descritiva sustenta a descrição dos padrões ortográficos do Português Brasileiro (doravante PB) utilizada na elaboração do questionário de coleta do conhecimento ortográfico dos sujeitos desta pesquisa. Por fim, a pesquisa de campo experimental envolve a coleta, controlada, de dados sobre o conhecimento ortográfico de alunos de uma escola pública em Belo Horizonte - Minas Gerais.

Esta pesquisa faz parte do projeto "Ensino de Língua Portuguesa pós-ensino remoto: aquisição da escrita, consciência linguística e ortografia", coordenado pela professora Dra. Daniela Mara Lima Oliveira Guimarães da Faculdade de Letras da Universidade Federal de Minas Gerais (UFMG). O projeto possui o registro no Comitê de Ética em Pesquisa (CEP) da UFMG sob o número 68266223.8.0000.5149. Todos os participantes maiores de idade ou responsáveis legais pelos menores, assinaram o Termo de Consentimento Livre e Esclarecido – TCLE – (Apêndice A e B) e os participantes menores de idade o Termo de Assentimento Livre e Esclarecido – TALE – (Apêndice C). O desenvolvimento desta tese está em conformidade com todos os termos éticos estabelecidos pelo CEP.

A seguir, detalharemos os procedimentos metodológicos adotados nesta pesquisa. Na primeira seção, há a descrição do estudo piloto, incluindo informações sobre os participantes, questionário, coleta e análise de dados, e por fim, as decisões metodológicas para a coleta de dados. Na segunda seção, serão descritos os procedimentos metodológicos utilizados para a coleta de dados, bem como os critérios para a análise de dados.

6.1 Estudo piloto: dos procedimentos metodológicos à coleta e à análise de dados

O teste piloto foi desenvolvido com o objetivo de testar o experimento e ajustar o que fosse necessário para a coleta de dados. Este estudo foi desenvolvido, remotamente, em janeiro de 2023 por meio do *Google Formulário*. A seguir, há a descrição do estudo piloto.

6.1.1 Participantes do estudo piloto

Os participantes desta fase da pesquisa foram alunos regularmente matriculados no 3º ano do Ensino Fundamental I, 6º e 9º anos do Ensino Fundamental II e na 3ª série do Ensino Médio (EM) em escolas públicas e privadas da cidade de Belo Horizonte/MG e região metropolitana. Este experimento piloto permitiu a participação de alunos de diversas escolas de cidade de Belo Horizonte/MG e de sua região metropolitana, como Sabará, Santa Luzia e Contagem. Os anos escolares, focos desta pesquisa, foram escolhidos pelos seguintes motivos:

- **3º ano do EFI (3A)**: segundo a Base Nacional Comum Curricular – BNCC (Brasil, 2017), o ensino formal das relações ortográficas se inicia a partir deste ano. O 1º e o 2º ano do EFI são destinados à alfabetização. Por isso, optamos, nesta pesquisa, por trabalhar com o 3º ano, por ser o primeiro ano do ensino da ortografia.

- **6º ano do EFII (6A)**: este ano é o primeiro ano do EFII. Além disso, os discentes passaram 3 anos sendo expostos ao ensino da escrita ortográfica. Por fim, segundo a BNCC (Brasil, 2017), espera-se que, neste ano, os alunos já escrevam palavras com relações ortográficas regulares e irregulares.

- **9º ano do EFII (9A)**: após três anos do 6º ano, o 9º ano é o encerramento do Ensino Fundamental e antecessor do Ensino Médio. De acordo com a BNCC (Brasil, 2017), neste ano, espera-se que os alunos dominem as relações ortográficas.

- **3ª Série do Ensino Médio (3EM)**: após três anos do 9º ano, o 3º ano do Ensino Médio é o encerramento da Educação Básica. Segundo a BNCC (Brasil, 2018), espera-se que o aluno, ao final da Educação Básica, tenha consciência e domínio da norma ortográfica. Por esperar um domínio consolidado da ortografia, este ano escolar foi escolhido como uma das variáveis da pesquisa.

O contato com os alunos ocorreu por meio do ambiente virtual, em redes sociais como *Facebook*, *Instagram* e *WhatsApp*. Além disso, a coleta de dados ocorreu via formulário do *Google Forms*. Participaram do teste piloto 23 alunos de escolas públicas e privadas de Belo Horizonte – MG e região metropolitana. Os estudantes foram organizados de acordo com o ano escolar, a saber: dois participantes do 3º ano do Ensino Fundamental (3A); cinco do 6º ano (6A); seis do 9º ano (9A) e 10 do 3º ano do Ensino Médio (3EM).

.

6.1.2 Estímulos: da elaboração de pseudopalavras à formulação do questionário

Nesta seção, os critérios para a elaboração dos padrões ortográficos são estabelecidos, assim como a classificação da frequência do padrão e a elaboração de pseudopalavras que compuseram os estímulos do estudo piloto.

Para a elaboração dos padrões ortográficos, foram consideradas as possibilidades fonotáticas do português brasileiro referente ao padrão silábico consoante e vogal (CV). Além disso, optou-se por trabalhar com os fonemas /s/, /ʃ/, /ʒ/. Tais fonemas podem ser representados por grafemas concorrentes irregulares no início de palavras diante de <e> e <i> e não há nenhuma regra que regule qual grafema utilizar para representar determinado fonema. Por exemplo, neste contexto de início de palavras, o fonema /s/ pode ser representado por <c> ou <s> (“cebola”, “semana”, “sistema”, “cicuta”); o /ʃ/ pelos grafemas <x> ou <ch> (“xepa”, “chefe”, “xícara”, “China”); e o fonema /ʒ/ por <g> ou <j> (“gente”, “jejum”, “girafa”, “jiló”). Portanto, os seguintes critérios foram considerados para a elaboração dos padrões ortográficos: padrão silábico CV; início de palavra; as consoantes <c> e <s>, <ch> e <x>, <j> e <g> diante das vogais <e> e <i>.

Após os padrões ortográficos estabelecidos, a frequência de tipo foi conferida no VOC (BECHARA, 2017). Para isso, por meio da busca “inicia com”, a frequência de tipo de cada padrão ortográfico foi conferida. Por exemplo, há 978 palavras que iniciam com o padrão <ge>, ou seja, este padrão se repete 978 vezes no PB. Para a definição de frequência, o número de ocorrências de palavras com determinado padrão ortográfico no corpus VOC (BECHARA, 2017) foi considerado. Além disso, para diferenciar o que é menor e maior frequência, foi considerado o número de ocorrência entre os grafemas concorrentes irregulares que formam o par na representação de determinado fonema. Por exemplo, o par de grafemas <se> e <ce> concorre para a representação do fonema /s/. O padrão ortográfico <ce> possui menor frequência (nº 1865) em relação ao padrão <se> (nº 3585), sendo este de maior frequência.

Após a definição dos padrões ortográficos e de sua frequência de tipo, as pseudopalavras que compuseram o questionário do teste piloto foram elaboradas. O uso de não-palavras, ou pseudopalavras, se justifica, pois, por meio deles, os sujeitos da pesquisa não puderam se basear na escrita de palavras reais previamente memorizadas (Altmiller; Treiman; Kessler, 2023). Para o estudo piloto, foram elaboradas 30 pseudopalavras, 10 de cada um dos fonemas estudados, e

três pseudopalavras com distratores. Todas estas pseudopalavras do teste piloto estão no Apêndice D.

6.1.3 Coleta de dados do estudo piloto

A coleta de dados do estudo piloto foi realizada remotamente por meio do *Google Formulário*. Os participantes da pesquisa receberam o link do formulário[18] via mensagem. No questionário, os sujeitos ouviram as pseudopalavras, gravadas no *Easy Voice Recorder* pela pesquisadora, e tiveram que digitar a letra que melhor representava o som escutado e que preenchesse a lacuna inicial da pseudopalavra. Como o foco da pesquisa é o contexto inicial da palavra, optou-se por deixar visível para o estudante o restante da pseudopalavra. Em cada questão, o aluno foi orientado a clicar no vídeo e ouvir a palavra. Após isso, ele foi questionado com a pergunta: "Qual a primeira letra (ou letras) da palavra que você ouviu anteriormente?". Na figura, a seguir, há um exemplo de questão com a pseudopalavra [siˈpapo].

Figura 7 – Exemplo de questão com a pseudopalavra [siˈpapo]

Fonte: Elaboração própria.

6.1.4 Análise e discussão de dados do teste piloto

Após a coleta, os dados foram transcritos, catalogados e organizados em planilha no *Excel*. Na Tabela 4 a seguir, há a descrição do ano escolar, na primeira coluna, o número de dados coletados por fonema em cada ano escolar nas colunas centrais, seguido pelo número total de dados coletados por ano escolar. Na parte inferior da tabela, é apresentado o número total de fonemas coletados.

[18] Link: https://forms.gle/uYtoP4SAL3W1XWwh7

Tabela 3 - Quantitativo de dados do teste piloto

Ano escolar	/s/	/ʃ/	/ʒ/	Total por ano escolar
3EF	20	19	20	59
6A	49	39	50	138
9A	53	49	54	156
3EM	100	99	98	297
Total por fonema	**222**	**206**	**222**	**650**

Fonte: *Corpus* da pesquisa.

O total de dados analisados foi de 650, como explícito na tabela anterior. Em relação ao fonema /s/, foram registrados 222 dados; para o fonema /ʃ/, foram 206; e para o /ʒ/, 222 dados. Em relação aos anos escolares, foram coletados 59 dados no 3EF, 138 dados no 6A, 156 dados no 9A e 297 dados no 3EM.

A partir da análise descritiva dos dados do estudo piloto, observamos indícios de que os três fatores, ano escolar, padrão ortográfico e frequência, poderiam influenciar a escrita de grafemas concorrentes irregulares, de forma isolada ou combinada. No entanto, para afirmar a significância desses indícios, foi necessário ajustar alguns pontos da metodologia para garantir a confiabilidade dos resultados. Portanto, na seção a seguir, há o levantamento das limitações do estudo piloto e os novos caminhos percorridos para a coleta dos dados.

6.1.5 Limitações do estudo piloto e novos caminhos para a coleta de dados

Durante o desenvolvimento do teste piloto, foram identificadas algumas limitações metodológicas. Assim, o objetivo desta seção é destacar essas limitações e descrever as estratégias utilizadas para reduzir seus impactos na coleta de dados. A seguir, são apresentadas cada uma dessas limitações e as estratégias adotadas para mitigá-las.

- **Tamanho do *Corpus*:** o baixo número de dados do teste piloto impossibilitou a aplicação de testes estatísticos, permitindo apenas a análise descritiva dos dados. Essa limitação pode ter ocorrido devido à dificuldade no contato com os participantes, o qual foi realizado apenas em ambiente virtual e ao possível desinteresse das pessoas em participar de pesquisas. Além disso, houve perda significativa de dados, porque os participantes não ouviram o áudio mais de uma vez e não compreenderam o som produzido. Na coleta, a aplicação da pesquisa foi realizada em uma escola e a pesquisadora coletou, presencialmente, os dados de cada participante, possibilitando

que eles ouvissem estímulo auditivo quantas vezes fosse necessário. Além disso, podemos controlar o número de participantes da coleta, em um total de 20 de cada ano escolar.

- **Tamanho do questionário:** Os participantes, principalmente os mais jovens do 3º, 6º e 9º anos do Ensino Fundamental, relataram cansaço e desânimo durante a resposta ao questionário. Para contornar esta limitação, estratégias lúdicas foram utilizadas na coleta de dados, como a nomeação de personagens *Digimon*[19] e desenhos animados (tais estratégias serão detalhadas na seção seguinte). Além disso, a pesquisadora sugeriu pausas aos sujeitos quando eles demonstravam desânimo e desinteresse.

- **Participantes de diferentes tipos de escolas:** por se tratar de um teste piloto, foi permitida a participação de alunos de escolas públicas e privadas de diferentes cidades da região metropolitana de Belo Horizonte – MG. Na coleta de dados, no entanto, participaram da pesquisa apenas alunos de uma escola estadual na região da Pampulha em Belo Horizonte – MG, visando à uniformidade do perfil dos participantes.

- **Impacto da frequência de ocorrência:** no estudo piloto, decidimos trabalhar apenas com a frequência de tipo e, por isso, elaboramos pseudopalavras. No entanto, após a revisão da literatura sobre a frequência de tipo e de ocorrência, foi perceptível a interação entre elas. Portanto, para a coleta de dados, optamos por trabalhar, além da frequência de tipo, também com a frequência de ocorrência.

- **Falta de aleatoriedade:** no teste piloto, cada participante preencheu o questionário com as pseudopalavras na mesma ordem. Na coleta de dados, garantimos a aleatoriedade e cada estudante recebeu um questionário com as palavras em ordem diferente.

- **Diferente percurso de aprendizagem do dígrafo <ch>:** no estudo piloto, o fonema /ʃ/ e sua relação com os grafemas <x> e <ch> foram considerados, sendo o último um dígrafo[20]. No entanto, o percurso de aprendizagem de dígrafos se difere de outros grafemas não dígrafos (Miranda; Pachalski; Richetti, 2023). Por isso, decidimos excluir o fonema /ʃ/ e seus grafemas <x> e <ch> e incluir o fonema /z/ e seus grafemas <s> e <z>.

[19] *Digimon* é uma franquia japonesa de desenho animado.
[20] Dígrafo é a junção de dois grafemas para representar um fonema. Exemplos de dígrafo: <rr> arroz, <ss>, pássaro.

6.2 Coleta de dados: dos procedimentos metodológicos à coleta e à análise de dados

Nesta seção, descrevemos os procedimentos metodológicos da coleta de dados. Há seções de descrição da escola, dos participantes, da elaboração do experimento e, por fim, a organização e análise dos dados.

6.2.1 Escola "Esperança"

Este experimento foi desenvolvido em uma escola estadual na região da Pampulha na cidade de Belo Horizonte – MG. Após tentativas, sem sucesso, de desenvolver a pesquisa em várias outras instituições, a escola "Esperança", pseudônimo criado pela pesquisadora, a acolheu muito bem, assim como a sua proposta de pesquisa. Os diretores, supervisores, professores e demais funcionários da escola receberam a pesquisadora de forma calorosa e a incluíram como integrante na equipe da escola. O nome "Esperança", além de manter o anonimato da instituição, foi escolhido para transmitir a mensagem de que há escolas públicas abertas ao desenvolvimento de pesquisas acadêmicas, e que há a esperança para uma maior integração entre universidade e escola básica.

A escola "Esperança" é uma das poucas escolas estaduais da cidade de Belo Horizonte que abarca todos os anos escolares da Educação Básica, como o Ensino Fundamental I e II, o Ensino Médio, e a Educação de Jovens e Adultos (EJA). A escola possui diversas salas de aula, laboratórios de ciências e informática, biblioteca, quadras esportivas e instalações com acessibilidade. A instituição é gerida por uma diretora e três vice-diretores, um para cada turno (manhã, tarde e noite). Segundo observação da pesquisadora, a escola "Esperança" é organizada, apresentando uma ótima estrutura física e uma boa relação entre todos os funcionários. Além disso, a instituição está situada em um dos bairros da Região da Pampulha na cidade de Belo Horizonte e atende moradores desta região, que são, em sua maioria, de classe média e média alta. De acordo com o Nível Socioeconômico (Inse) do Sistema de Avaliação da Educação Básica (Saeb) de 2021 (Brasil, 2023), os alunos da escola "Esperança" foram classificados em Nível VI. Isso significa que esses discentes estão em uma condição socioeconômica superior à média nacional, mas ainda não estão entre os índices mais altos. Esse nível indica que as famílias dos estudantes têm uma estrutura financeira relativamente estável e acesso a diversos bens e serviços.

O Índice de Desenvolvimento da Educação Básica (Ideb), cuja função é medir a qualidade da educação brasileira em uma escola de 0 a 10, estabeleceu a meta de alcançar 6

pontos até 2022. Ao analisar os resultados do Ideb (Brasil, 2024) do ano de 2023 em relação às escolas estaduais, a instituição “Esperança” está acima da média nacional em cada segmento educacional, Ensino Fundamenta I e II e Ensino Médio, como destacado na tabela a seguir.

Tabela 4 - Dados do Ideb de 2023 da escola “Esperança”

Segmento educacional	Média Nacional	Média da escola “Esperança”
Ensino Fundamental I	6,0	7,9
Ensino Fundamental II	4,9	5,6
Ensino Médio	4,1	4,2

Fonte: Brasil (2024).

Embora a escola esteja acima da média nacional em todos os seguimentos educacionais, apenas o resultado do EFI está acima da meta educacional estabelecida para 2022. Além disso, esta média cai ao chegar ao final da educação básica, o Ensino Médio. Portanto, a escola “Esperança” demonstra um desempenho excelente no EFI, um desempenho satisfatório no EFII e enfrenta desafios no Ensino Médio.

Para o desenvolvimento da pesquisa, os vice-diretores do turno da manhã e da tarde assinaram a Carta de Anuência (Apêndice E) concedendo permissão para a realização da pesquisa em ambos os turnos. Os vice-diretores e as supervisoras, gentilmente, auxiliaram na coleta dos termos de consentimentos dos responsáveis e dos termos de assentimento dos alunos menores de idade. A seguir, há a descrição dos participantes da pesquisa.

6.2.1.1 Participantes

Todos os participantes da coleta de dados são alunos regulares da escola “Esperança”. Assim como no estudo piloto, nesta coleta os sujeitos são alunos do 3º ano do Ensino Fundamental I (3A), do 6º (6A) e 9º anos (9A) do Ensino Fundamental II e da 3ª série (3EM) do Ensino Médio. A justificativa para a escolha desses anos escolares foi descrita na seção do “Estudo Piloto”. A pedagoga e/ou o vice-diretor visitaram cada sala para apresentar a pesquisadora aos alunos e convidá-los a participar da pesquisa. Inicialmente, foi informado que até 20 discentes poderiam participar. A seleção ocorreu de forma aleatória entre os interessados em cada turma. Como havia várias turmas para um único ano escolar, os participantes selecionados pertenciam a diferentes turmas. Por exemplo, são três turmas de 6º ano, quatro turmas do 9º ano.

Na Tabela 5, a seguir, há o número (*Nº*) de participantes por ano escolar, e, na última linha, o total de sujeitos.

Tabela 5 – Número de participantes da coleta de dados por ano escolar

Ano escolar	***Número de participantes***
3º ano do EF (3A)	20
6º ano do EF (6A)	20
9º ano do EF (9A)	20
3ª série do EM (3EM)	20
Total	**80**

Fonte: *Corpus* da pesquisa.

A Tabela 5 evidencia que 80 discentes participaram do experimento, sendo 20 alunos de cada ano escolar. Para a participação na pesquisa, as assinaturas, dos discentes e de seus responsáveis, nos termos de consentimento e assentimento foram coletadas. Apenas participaram da pesquisa os alunos que trouxeram os termos assinados pelos responsáveis.

A maioria dos alunos demostrou entusiasmo para participar da pesquisa, sendo que no 3º ano do EF (3A) houve disputa para saber quem seria o próximo a ser chamado para participar. Infelizmente, alguns pais não permitiram que seus filhos participassem, o que deixou os alunos despontados ao não serem chamados pela pesquisadora. Os alunos do 6º ano (6A) e da 3ª série do EM (3EM) também demonstraram interesse e entusiasmo em participar da pesquisa. Os discentes do 3EM, especificamente, mostraram-se curiosos para saber o objetivo da pesquisa. Apenas os alunos do 9º ano do EF (9A) que demonstraram desinteresse pela pesquisa, mas aceitaram a participar.

6.2.2 Experimento

O experimento foi organizado em duas partes o "Tarefa 1 – Pseudopalavras" (Apêndice F) e o segundo, "Tarefa 2 – Palavras reais" (Apêndice G). Nas seções a seguir, há a descrição de como cada uma dessas partes foi elaborada, como compuseram o questionário.

6.2.2.1 Tarefa 1: Pseudopalavras

Para a elaboração dos estímulos das pseudopalavras, a frequência de tipo dos grafemas foi verificada na plataforma VOC (Bechara, 2017). No Capítulo 4 – "Frequência em evidência", foi apresentada a justificativa e a explicação do porquê do uso dessa plataforma.

Além disso, como argumentado no mesmo Capítulo quatro, há uma falta de critérios para a definir o que é de maior ou menor frequência nos trabalhos que consideram a frequência como variável de análise. Para suprir esta lacuna, realizamos o teste Qui – Quadrado de Pearson para que seja possível afirmar, estatisticamente, o que é de maior e menor frequência de tipo. O teste Qui – Quadrado de Pearson é utilizado para medir a independência entre variáveis (Franke; Ho; Cristie, 2012). Nesta tese, mediremos se a frequência de tipo de um grafema é, estatisticamente, diferente do seu grafema concorrente. Por exemplo, observaremos se a frequência de tipo do grafema <s>, no contexto início de palavra diante de <e> e <i>, é maior do que o grafema <c> no mesmo contexto para representar o fonema /s/. Consideramos, nesta tese, o p-valor 0.05.

A Tabela 6, a seguir, retoma o objeto de estudo desta pesquisa, assim como a frequência de tipo, a proporção e o resultado do teste Qui-quadrado. Na primeira coluna, há os fonemas, seguidos pelo contexto de ocorrência observado nesta tese. Na coluna "Grafema", há as possibilidades de grafema que podem ser utilizados para representar o fonema em questão. Além disso, na coluna "*N*-grafema", há a frequência de tipo, ou seja, o número de vezes em que determinado grafema ocorreu, no contexto estudado, segundo os dados da plataforma VOC. Por fim, na última coluna, há o resultado do teste Qui – Quadrado de Pearson, utilizado para definição de qual dos grafemas concorrentes era o de maior ou menor frequência.

Tabela 6 - Frequência de tipo de grafemas do PB

Fonema	Contexto	Grafema	*N*-grafema	X^2
/s/	Início de palavra diante de <e> e <i>	<c>	3682	X-squared = 68677, df = 3, p-value < 2.2e-16
		<s>	5666	
/ʒ/	Diante de <e> e <i> em início de palavra	<g>	1476	X-squared = 1009.8, df = 3, p-value < 2.2e-16
		<j>	444	
/z/	Entre vogais	<s>	110597	X-squared = 68677, df = 3, p-value < 2.2e-16
		<z>	92033	

Fonte: Elaborado a partir de Bechara (2017).

Na Tabela 6, o fonema /s/ ocorre 9348 vezes em contexto inicial de palavra diante de [e] e [i]. Este fonema, nesse mesmo contexto, pode ser representado por dois grafemas, o <s> e o <c>. O grafema <s> tem frequência de 5666 e o <c> de 3682, ou seja, o primeiro ocorre 5666 vezes e o segundo 3682 vezes no contexto descrito para representarem o fonema /s/ (Bechara, 2017). O teste Qui-quadrado, como podemos observar na última coluna da Tabela 6,

indicou que há diferença significativa entre a frequência de <s> e <c> no contexto considerado. Portanto, podemos afirmar que <s> é de maior frequência de tipo e o <c> de menor.

Na segunda linha da tabela, há o fonema /ʒ/ pode ser representado por <g> e <j> diante de <e> e <i>. É importante ressaltar que, também, estamos considerando o contexto inicial de palavra nesta relação fonema-grafema. O grafema <g> ocorre 1476 vezes e o <j> 444, ambos no mesmo contexto (Bechara, 2017). O teste Qui-quadrado, como podemos observar na última coluna da Tabela 6, indicou que há diferença significativa entre a frequência de <g> e <j> no contexto considerado. Portanto, podemos afirmar que <g> é de maior frequência de tipo e o <j> de menor.

Na terceira linha da Tabela 6, o fonema /z/ pode ser representado pelos grafemas <s> e <z> em contexto intervocálico, sendo o primeiro com frequência de 110597 e o segundo de 92033. O resultado do Qui-quadrado, como podemos observar na última coluna da Tabela 6, indicou diferença entre a frequência de <s> e <z> no contexto intervocálico. Portanto, podemos afirmar que <s> é de maior frequência de tipo e o <z> de menor. Na Tabela 7, a seguir, estão as pseudopalavras elaboradas para este experimento.

Tabela 7 – Transcrição fonética das pseudopalavras elaboradas para a realização do experimento

Fonema	**Pseudopalavras**				**Total por fonema**
/s/	['sipə]	['silʊ]	['sibɪ]	['sikə]	8
	['sevʊ]	['sehɪ]	['setʊ]	['senɪ]	
/ʒ/	['ʒidʃɪ]	['ʒikʊ]	['ʒimə]	['ʒivɪ]	8
	['ʒefʊ]	['ʒetʊ]	['ʒepʊ]	['ʒedʊ]	
/z/	['tezə]	['bezə]	['kezə]	['fezə]	8
	['dʒizə]	['hizə]	['kizə]	['nizə]	
Total de pseudoplavras					**24**

Fonte: Elaboração própria.

As pseudopalavras foram elaboradas a partir dos seguintes critérios:

- Pseudopalavras dissílabas;
- Paroxítonas;
- Sílaba simples, consoante – vogal;
- Som [s] e [ʒ] sempre no início de palavra diante de [e] e [i];

- Som [z] em contexto intervocálico.

6.2.2.2 Tarefa 2: Palavras reais

Nesta seção, há a descrição de como as palavras reais do experimento foram selecionadas. Devido à diferente faixa etária dos participantes desta pesquisa, dos oito aos 18 anos de idade, e, consequentemente, às suas diferentes vivências, as palavras reais deste experimento foram selecionadas, além de levar em consideração os fonemas e grafemas objetos de estudo desta tese, a partir de palavras de menor e maior ocorrência em dois *corpora*, o Léxico do Português Brasileiro (LexPorBR) e o LexPor - Infantil. No Capítulo 4 – "Frequência em evidência", as justificativas e a explicação do porquê do uso dessas plataformas são discutidas. Para a seleção das palavras, os seguintes critérios foram considerados:

- Frequência de ocorrência similar em dois *corpora*, um infantil e outro não-infantil;
- Grafemas <c> e <s> diante de <e> e <i>;
- Grafemas <g> e <j> diante de <e> e <i>;
- Grafemas <s> e <z> entre vogais;
- Mesma posição na palavra, início, meio ou fim;
- Mesma posição quanto à tonicidade;
- Mesmo número de sílabas nos pares de palavras.

Na Tabela 8, a seguir, há a lista de palavras selecionadas para este experimento. Na primeira coluna da tabela, há os fonemas (/s/, /ʒ/ e /z/) seguidos por seus grafemas (<s>, <c>, <g>, <j>, <s>, <z>) e palavras que contêm estes. Os grafemas observados nas palavras estão sublinhados e em negrito. Nas colunas "Frequência - LexPorBR Infantil" e "Frequência – LexPorBR", há a ocorrência das palavras no corpus infantil e no não-infantil, respectivamente. A coluna "Tonicidade" refere-se à posição que o grafema observado assume na palavra. Por exemplo, na palavra "cidade", o grafema <c> assume a posição pretônica. E na última coluna, o nº de sílabas de cada palavra é apresentado.

Tabela 8 – Palavras reais utilizadas no experimento, sua frequência nos *corpora* LexPorInfantil e LexPorBR e classificação quanto à tonicidade e ao número de sílabas

Fonema	Grafema	Palavra	Ocorência LexPorBR Infantil	Ocorrência LexPorBR	Tonicidade	Nº de sílaba
/s/	<c>	**c**idade	41918	16093	pretônica	3
		cicuta	63	18	pretônica	3
		certo	312551	4668	tônica	2
		cerne	38	76	tônica	2
	<s>	**s**emana	46151	14927	pretônica	3
		sequela	49	21	pretônica	3
		sistema	10879	10723	pretônica	3
		singelo	26	40	pretônica	3
/ʒ/	<g>	**g**eneral	8051	1849	pretônica	3
		gestual	19	33	pretônica	3
		gigante	5847	206	pretônica	3
		gincana	57	26	pretônica	3
	<j>	su**j**eito	3667	1127	tônica	3
		de**j**eto	13	5	tônica	3
		jipe	552	114	tônica*	2
		jiló	1	12	pretônica*	2
/z/	<z>	surpre**s**a	17953	1220	postônica	3
		turque**s**a	123	14	postônica	3
		bele**z**a	16301	1162	postônica	3
		leve**z**a	107	115	postônica	3
	<s>	ami**z**ade	5813	608	tônica*	4
		oje**r**iza	2	25	postônica*	4
		cami**s**a	8610	1207	postônica	3
		divi**s**a	72	170	postônica	3

Fonte: Estivalet (2019) e Estivalet e colaboradores (2023).

Na Tabela 8, os pares de palavras da mesma cor (cinza e branco) foram estabelecidos quanto à frequência e classificados quanto à tonicidade e número de sílabas. Por exemplo, no par “cidade” e “cicuta”, o grafema <c> está na mesma posição, início de palavra, pretônica, e há o mesmo número de sílabas (três) nas duas palavras. Além disso, a frequência de ocorrência dos dois vocábulos é sempre oposta nos dois *corpora*. Por exemplo, “cidade” tem uma maior frequência no LexPor Infantil (41918) e no LexPor não-infantil (16093) em relação à menor frequência da palavra “cicuta” no LexPor Infantil (63) e no LexPor não-infantil (18). Esta mesma lógica é aplicada em todos os outros pares de palavras, como “certo” e “cerne”; “semana” e “sequela”; “sistema” e “singelo”; “general” e “gestual”; “gigante” e “gincana”; “sujeito” e “dejeto”; “jipe” e “jiló”; “surpresa” e “turquesa”; “beleza” e “leveza”; “amizade” e “ojeriza”; “camisa” e “divisa”. Apenas nos pares destacados pelos asteriscos, “jipe” e “jiló” e “amizade” e “ojeriza”, não foi possível controlar a tonicidade. No entanto, como o grafema observado ocorre no mesmo contexto, ou seja, início de palavra e diante de <i> no primeiro par, e entre as vogais <i> e <a>, optamos por considerar estes pares.

Como discutido no Capítulo 4 – "Frequência em evidência", há uma falta de parâmetros em relação ao que é palavra frequente e palavras infrequente no português. Para suprir esta lacuna, o teste Qui – Quadrado de Person, indicado para medir independência entre variáveis (Franke; Ho; Cristie, 2012), foi utilizado nesta tese. Dessa forma, ainda que as palavras tenham sido selecionadas a partir do número de vezes que ela ocorre no português brasileiro, a classificação dos vocábulos como de maior ou menor ocorrência nesta tese foi realizada a partir de resultados significativos no teste Qui – Quadrado de Person ($p<0,05$) em ambos os *Corpora*.

Inicialmente, as palavras foram separadas em dois grupos (A e B) a partir do número de vezes em que ela apareceu em cada um dos dois *Corpus*, sendo o grupo A composto pelas palavras com maior ocorrência, e o grupo B, pelas palavras com menor ocorrência. Em seguida, aplicou-se o Teste Qui – quadrado de Person entre os grupos de palavras, com o objetivo de verificar a diferença entre a sua ocorrência ($p<0,05$). Na Tabela 9, a seguir, há a aplicação do Teste Qui-Quadrado.

Tabela 9 – Aplicação do Teste Qui – quadrado na ocorrência de palavras do LexPorBR - Infantil

Fonema	**Grupo A**		**Grupo B**		***X²***
	Palavra	*Ocorrência*	*Palavra*	*Ocorrência*	
/s/	cidade	41918	cicuta	63	X^2-squared = 1564515, df = 7, p-value < 2.2e-16
	certo	312551	cerne	38	
	semana	46151	sequela	49	
	sistema	10879	singelo	26	
/ʒ/	general	8051	gestual	19	X^2-squared = 31339, df = 7, p-value < 2.2e-16
	gigante	5847	gincana	57	
	sujeito	3667	dejeto	13	
	Jipe	552	jiló	1	
/z/	surpresa	17953	turquesa	123	X^2-squared = 64694, df = 7, p-value < 2.2e-16
	beleza	16301	leveza	107	
	amizade	5813	ojeriza	2	
	camisa	8610	divisa	72	

Fonte: Elaborada a partir do LexPorBR – Infantil (Estivalet, 2019).

Na primeira coluna da tabela, há os fonemas. Na coluna nomeada Grupo A, há as palavras com o maior número de vezes que apareceram no corpus (Ocorrência). No Grupo B, há as palavras de menor número de ocorrência. Na última coluna (X^2), há o resultado do Teste Qui – quadrado.

Três testes do Qui – quadrado foram realizados, um para cada um dos conjuntos de palavras a partir do fonema, havendo diferença significativa entre os Grupos A e B. Dessa

forma, a hipótese nula foi rejeitada e a hipótese alternativa confirmada de que uma diferença significativa entre os grupos com o Grupo A tendo uma frequência maior do que o Grupo B. Portanto, podemos afirmar que as palavras selecionadas no *corpus* LexPorBR – Infantil (Estivalet, 2019) que compõem o Grupo A e o Grupo B se diferem estatisticamente.

No entanto, como não trabalhamos apenas com o *corpus* infantil, realizamos os mesmos testes nas palavras selecionadas no LexPorPB não- infantil (Estivalet *et al*., 2023). Na Tabela 10, a seguir, há as palavras e o número de ocorrência de cada uma delas. No Grupo A, há palavras com um maior número de ocorrência, e, no Grupo B, com menor número. Na última coluna, há o resultado do teste.

Tabela 10 - Aplicação do Teste Qui – quadrado na ocorrência de palavras do LexPorPB não-infantil

Fonema	**Grupo A**		**Grupo B**		**X^2**
	Palavra	*Ocorrência*	*Palavra*	*Ocorrência*	
/s/	cidade	16093	cicuta	18	X^2-squared = 59706, df = 7, p-value < 2.2e-16
	certo	4668	cerne	76	
	semana	14927	sequela	21	
	sistema	10723	singelo	40	
/ʒ/	general	1849	gestual	33	X^2-squared = 7936, df = 7, p-value < 2.2e-16
	gigante	206	gincana	26	
	sujeito	1127	dejeto	5	
	jipe	114	jiló	1	
/z/	surpresa	1220	turquesa	14	X^2-squared = 3810.1, df = 7, p-value < 2.2e-16
	beleza	1162	leveza	115	
	amizade	608	ojeriza	25	
	camisa	1207	divisa	170	

Fonte: Elaborado a partir de LexPorPB (Estivalet *et al.*, 2023).

Assim como no LexPorBR – Infantil, os resultados dos testes nos grupos de palavras do LexPorPB, não-infantil, indicaram que há diferença significativa entre o Grupo A e o Grupo B. Portanto, as palavras selecionadas nos dois *corpora* são estatisticamente diferentes, havendo palavras de maior e de menor frequência de ocorrência.

Na Tabela 11 a seguir, estão as palavras classificadas quanto à frequência de ocorrência. Na primeira coluna, há os fonemas foco deste estudo, seguidos pelas palavras que os contêm. O grafema que representa o fonema da coluna anterior está em negrito e sublinhado. Na próxima coluna, há a classificação da frequência de ocorrência das palavras em cada linha. Na última

coluna, há o número de palavras reais por fonema, sendo oito em cada um deles. Por fim, na última linha, há o total de palavras reais do experimento, ou seja, 24 itens.

Tabela 11 – Palavras reais do experimento

Fonema	Palavras reais				Frequência de ocorrência	Total por fonema
/s/	**c**idade	**c**erto	**s**emana	**s**istema	maior	8
	cicuta	**c**erne	**s**equela	**s**ingelo	menor	
/ʒ/	**g**eneral	**g**igante	**j**ipe	su**j**eito	maior	8
	gestual	**g**incana	**j**iló	de**j**eto	menor	
/z/	surpre**s**a	bele**z**a	ami**z**ade	**c**ami**s**a	maior	8
	turque**s**a	leve**z**a	ojeri**z**a	divi**s**a	menor	
					Total de palavras	**24**

Fonte: Elaboração própria.

6.2.2.3 Organização dos estímulos

Os estímulos do experimento foram divididos em duas partes: a primeira "Tarefa 1 – Pseudopalavras" (Apêndice F) e a segunda, "Tarefa 2 – Palavras reais" (Apêndice G). Em cada parte, há 32 estímulos, 24 alvos e 8 distratores. Estes equivalem a 1/3 do estímulo alvo. Portanto, cada participante teve contato com 64 estímulos no total (os estímulos estão disponíveis nos Apêndices F e G).

A ordem das pseudopalavras e palavras reais foi aleatória[21] para cada um dos sujeitos. Para que todos estes recebessem o mesmo estímulo auditivo, a pesquisadora gravou cada uma das palavras e pseudopalavras de acordo com o dialeto mineiro. Além disso, cada estímulo auditivo continha um estímulo visual. Nas pseudopalavras, imagens de personagens de desenhos animados, como "Rugrats" e "A hora do recreio", foram relacionadas aos distratores, e as criaturas do desenho "Digimon" foram relacionados aos estímulos alvos. As imagens e os sons gravados foram dispostos em apresentação do *PowerPoint*, um excelente recurso para a organização de imagem e áudio, além de permitir a aleatoriedade dos estímulos.

[21] Um agradecimento à Cecília Toledo que disponibilizou e orientou o uso do *script* para aleatoriedade de slides do *PowerPoint*.

6.2.3 Coleta de dados

Após a assinatura do TCLE pelos responsáveis e do TALE, a coleta de dados foi iniciada. Para auxiliar a coleta de dados, um documento com os procedimentos de coleta foi desenvolvido e está no Apêndice H.

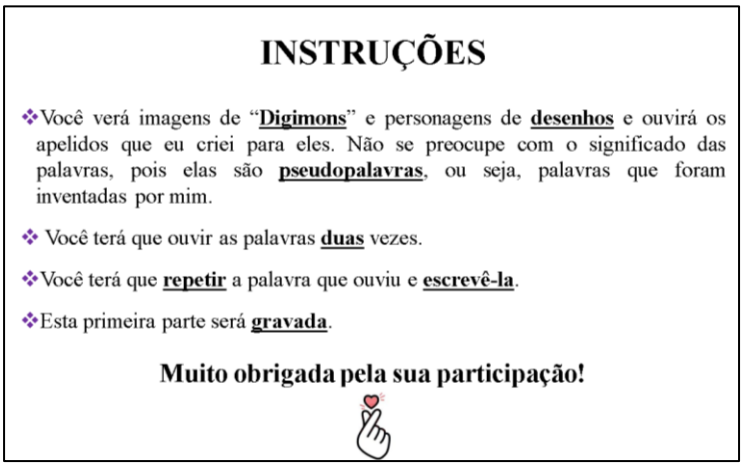
O experimento foi realizado no período de sete dias, nos turnos da manhã e da tarde, no mês de setembro de 2023. Os materiais utilizados para a coleta foram: computador, fone de ouvido, celular com gravador de voz, lápis, caneta, papel, todos disponibilizados pela pesquisadora, mesa e cadeira disponibilizados pela escola "Esperança". A coleta de dados foi realizada em um dos laboratórios de ciências da escola. Este local foi ideal, pois havia cadeiras e bancadas para os alunos apoiarem os braços e escreverem, além de ser um ambiente calmo e silencioso, o que permitiu que os participantes ouvissem os áudios do questionário nitidamente. A coleta de dados foi individualmente com cada participante e foi realizada em duas partes, no mesmo dia, a primeira "Tarefa 1 – Pseudopalavras" e a segunda "Tarefa 2 – Palavras reais".

Cada aluno foi chamado à sala e conduzido, pela pesquisadora, ao laboratório de ciências. No caminho para o laboratório, a pesquisadora conversava com o aluno sobre como era estudar naquela escola e se ele já havia participado de alguma pesquisa anteriormente. A estratégia da conversa foi utilizada para criar um vínculo com o estudante e fazê-lo se sentir mais confortável para participar das atividades da pesquisa.

Ao chegarem ao laboratório, a pesquisadora explicava que, inicialmente, o aluno ouviria palavras que não existiam no português, mas que eram apelidos criados para os *Digimons* e desenhos animados de que a pesquisadora mais gostava. Além disso, a pesquisadora explicava que não havia respostas certas e erradas, que ela gostaria de conhecer o que o estudante pensava e sabia sobre a escrita de algumas palavras. Foi também orientado que a primeira parte da atividade, "Tarefa 1 – Pseudopalavras", seria gravada.

Cada participante recebeu um lápis ou caneta e uma folha de resposta (Apêndice I), contendo espaço para nome, ano escolar, idade, e linhas para a escrita de cada estímulo sonoro. A pesquisadora lia, com o estudante, as orientações dispostas no *PowerPoint*, como exemplificado na figura a seguir.

Figura 8 - Instruções para a realização do experimento



Fonte: Elaboração própria.

Após a leitura das orientações, o participante era direcionado a colocar os fones de ouvido para realizar o treinamento e testar se o volume estava bom. As orientações da Tarefa 2 – Palavras reais também se assemelham às apresentadas na imagem. Na figura, a seguir, há um exemplo do treinamento.

Figura 9 - Exemplo do treinamento para o início do experimento

Fonte: Elaboração própria.

A Figura 8 apresenta apenas uma ilustração do treinamento com a transcrição fonética do áudio que o aluno ouviu. Na coleta, o participante não recebeu nenhum estímulo de escrita. Após o treinamento, a pesquisadora elucidava as dúvidas, caso surgissem, e dava prosseguimento ao experimento. Foi permitido ao aluno ouvir o áudio quantas vezes fosse necessário. No entanto, a pesquisadora não esclareceu dúvidas sobre a escrita das palavras.

Além disso, durante o experimento com as pseudopalavras, os sujeitos eram questionados sobre suas escolhas em relação à escrita de determinada letra. Por exemplo, a pesquisadora realizou perguntas como: *Por que você escreveu esta palavra com a letra z?* Estas respostas foram gravadas pelo celular da pesquisadora. Estas perguntas foram realizadas sempre nas mesmas pseudopalavras com alunos de diferentes anos escolares. As pseudopalavras foram escolhidas por fonema e aleatoriamente, tais como: ['silʊ], ['sevʊ], ['ʒepʊ], ['ʒikʊ], ['bezə], ['nizə], ['tezə], ['dʒizə]. Considerou-se realizar as perguntas apenas ao final do experimento, para evitar possíveis influências nas respostas. Contudo, isso poderia fazer com que o aluno esquecesse os motivos que o levaram àquela escrita. Por esse motivo, optamos por limitar o número de perguntas e aplicá-las no momento da escrita de apenas oito estímulos, buscando captar diretamente o que o aluno pensou, sem influenciar suas respostas posteriores.

Ademais, é importante ressaltar que a pesquisadora realizou as perguntas apenas para participantes que se demonstraram à vontade durante a pesquisa. Alguns alunos estavam tímidos e com receio de falar algo de errado, mesmo que a pesquisadora tenha assegurado, no início da pesquisa, que ela gostaria de conhecer o que o aluno pensa sobre a escrita, que não caberia, no momento, certo e errado. A maioria dos discentes do 3º ano do Ensino Fundamental I (EFI) estavam tímidos e, por isso, a pesquisadora optou por não os questionar sobre a letra escrita. Devido a este contexto, não há uma variabilidade de dados para o 3º ano do EFI. Além disso, os alunos do 9º ano, como relatado no Capítulo da metodologia, não apresentaram interesse na pesquisa, por isso, ao serem questionados sobre o motivo que os levaram a escrever determinada letra, as respostas foram “sei lá”. Diante desse cenário, a pesquisadora também respeitou o desinteresse dos alunos e não realizou as perguntas. Por isso, não há nenhum dado de transcrição de gravações do 9º ano do Ensino Fundamental II (EFII).

Após finalizada a primeira parte do experimento, a pesquisadora perguntava se estava tudo bem, se os participantes gostariam de alguma pausa e se a pesquisa poderia continuar para a segunda parte, com as palavras reais. Caso fosse necessária uma pausa para descanso, a pesquisadora permitia a saída do estudante para beber água ou ir ao banheiro e retornar ao laboratório para finalizar o experimento. Ao final do experimento, a pesquisadora agradecia a participação, acompanhava o aluno à sala de aula e chamava, individualmente, o próximo participante.

6.2.4 Análise de dados

Após a coleta, os dados foram organizados e catalogados. Os dados de escrita foram organizados em uma planilha do *Excel* e as gravações foram transcritas em arquivo do Word. A análise de dados foi realizada em duas partes, a primeira consistiu na análise quantitativa e a segunda, na análise qualitativa.

6.2.4.1 Parte 1: Análise quantitativa

Para a parte quantitativa, a planilha do *Excel* foi organizada de acordo com as variáveis da pesquisa e convertida para o formato .csv. As análises estatísticas foram conduzidas no *software* R Core Team (2023). O *R* é um ambiente de *software* livre para a computação estatística e gráfica, que tem sido amplamente utilizado em pesquisas linguísticas (Oushiro, 2017; Winter, 2019; Gries, 2019). Os *scripts* desenvolvidos no *software* R (2023) estão disponíveis no Apêndice J – "Scripts das análises estatísticas no R".

Para facilitar a compreensão da organização dos dados na tabela do *Excel*, apresentamos no quadro a seguir a descrição de cada uma das colunas da tabela. Os acentos gráficos foram retirados propositadamente e os símbolos dos fonemas não foram utilizados para facilitar a leitura dos dados no R (2023). Cada linha do Quadro 12 representa uma coluna da tabela do *Excel*. Na primeira coluna do quadro, há as variáveis e na segunda coluna a descrição destas.

Quadro 12 – Descrição das colunas da tabela de dados no Excel

Variável	Descrição
participante	Identificação do participante por código. O código "GV3A" indica um aluno "GV" que está no 3º ano do EF
ano	Identificação do ano escolar, 3A, 6A, 9A e EM3
estimulo	Palavras reais e pseudopalavras produzidas pela pesquisadora
focorrenciacat	Frequência de ocorrência de palavras reais. Níveis da variável: maior/menor
fonema	Fonemas /s/, /ʒ/ e /z/ são representados como sa, ja, za, respectivamente
grafema	Grafemas <s>, <c>, <g>, <j>, <z> e <s> para as palavras reais e as combinações <s-c>, <g-j> e <z-s> para as pseudopalavras
ftipografema	Número absoluto da frequência de tipo
tipo	Identificação de pseudoplavra (pseudo) ou palavra real (real)

Variável	Descrição
escrita	A escrita do participante em cada palavra e pseudopalavra no quesitonário
letra	A letra que o participante escreveu no questionário
ftipoletra	Frequência de tipo da letra que foi escrita
certoerrado	Se a escrita da palavra real está certa ou errada de acordo com o Acordo Ortográfico

Fonte: Elaboração própria.

Ooptamos por, nesta tese, desenvolver as análises estatísticas em duas etapas. Na primeira etapa, analisamos os dados da Tarefa 1 – Pseudopalavras. Já na segunda etapa, realizamos as análises do Tarefa 2 – Palavras reais. Essa escolha se justifica pois, há uma diferença no processamento de palavras reais para as inventadas. Pesquisas apontam que as palavras reais podem ser reconhecidas pelos alunos de forma mais rápida do que as pseudopalavras (não-palavras) (Cortez, 2018; Bonini; Keske-Soares, 2018). Isso ocorre, pois, o processamento psicolinguístico de palavras reais requer competências linguísticas diferentes de pseudopalavras (Cortez, 2018). Além disso, o trabalho com estas pode "eliminar a interferência da familiaridade" e diminuir o "risco de confusão com outras palavras" (Bonini; Keske-Soares, 2018, p.5). Portanto, a forma como a memória é acessada no momento de uma pesquisa linguística pode ser diferente nas palavras e nas pseudopalavras.

Para o desenvolvimento de uma análise estatística, é importante definir as variáveis respostas (dependentes) e as variáveis preditoras (independentes). A variável resposta (ou dependente), de acordo com Godoy e Nunes (p.31, em preparação), é a "característica medida ou observada pelo pesquisador". Já a variável preditora (ou independente) "explica, de alguma forma, a variável resposta" (Godoy; Nunes, p.31, em preparação). Nesta tese, há duas variáveis respostas, uma para a Tarefa 1 – Pseudopalavras e outra para o Tarefa 2 – palavras reais. Na Tarefa 1, a variável resposta é a frequência de tipo (qualitativa categórica) do grafema escrito pelos participantes. Já na Tarefa 2, é a variável certo-errado (qualitativa categórica), ou seja, se os estudantes escreveram de acordo com as regras ortográficas. Já as preditoras, nesta pesquisa, são o ano escolar (qualitativa ordinal), a frequência de ocorrência (qualitativa categórica) e a relação fonema-grafema (qualitativa categórica). No quadro, a seguir, há a sistematização das variáveis e seus níveis.

Quadro 13 – Variáveis resposta e preditoras desta tese

Variável resposta (dependente)	
Tarefa 1	**Tarefa 2**
Frequência de tipo	Erro-acerto ortográfico
Variável preditora (independente)	**Níveis**
Ano escolar	- Nível 1: 3º ano do Ensino Fundamental I (3A) - Nível 2: 6º ano do Ensino Fundamental II (6A) - Nível 3: 9º ano do Ensino Fundamental II (9A) - Nível 4: 3ª Série do Ensino Médio (EM3)
Frequência de ocorrência	- Nível 1: maior frequência - Nível 2: menor frequência
Relação fonema-grafema	- Nível 1: Fonema /s/ e o par de grafema <c - s> - Nível 2: Fonema / ʒ/ e o par de grafema <g - j> - Nível 3: Fonema /z/ e o par de grafema <s - z>
Frequência de tipo	- Nível 1: menor frequência de tipo - Nível 2: maior frequência de tipo
Certo ou errado	-Nível 1: certo - Nível 2: errado

Fonte: Elaboração própria.

Na Tabela 13 a seguir, há o número total de dados coletados, considerando as pseudopalavras e palavras reais, de acordo com o fonema e o ano escolar.

Tabela 12 - Número de dados da coleta

PSEUDOPALAVRAS					
Fonema	**3º ano EFI**	**6º ano EFII**	**9º ano EFII**	**3ª série EM**	**Subtotal**
/s/	160	160	160	160	640
/ʒ/	160	160	160	160	640
/z/	160	160	160	160	640
Subtotal	**480**	**480**	**480**	**480**	**1920**
PALAVRAS REAIS					
/s/	160	160	160	160	640
/ʒ/	160	160	160	160	640
/z/	160	160	160	160	640
Subtotal	**480**	**480**	**480**	**480**	**1920**
Total	**960**	**960**	**960**	**960**	**3840**

Fonte: *Corpus* da pesquisa.

O número total de dados da pesquisa é 3840, dentre eles 1920 são de pseudopalavras e 1920 de palavras reais. Há 480 dados por ano escolar e 640 por fonema nas pseudopalavras e nas palavras reais. Ao observar o fonema e o ano escolar, há 160 dados de cada um dos fonemas em cada um dos anos escolares.

Como a resposta foi de livre escolha, ou seja, o aluno escreveu o que ele compreendeu do áudio escutado, optamos por não excluir dados que fugiram do esperado. Por exemplo, ao escutar uma palavra que inicia com [s], há as duas possibilidades de escrita, o grafema <s> ou o <c>, no entanto, o aluno escreveu com <f> ou <v>. Estes dados também foram considerados.

6.2.4.2 Parte 2: Análise qualitativa

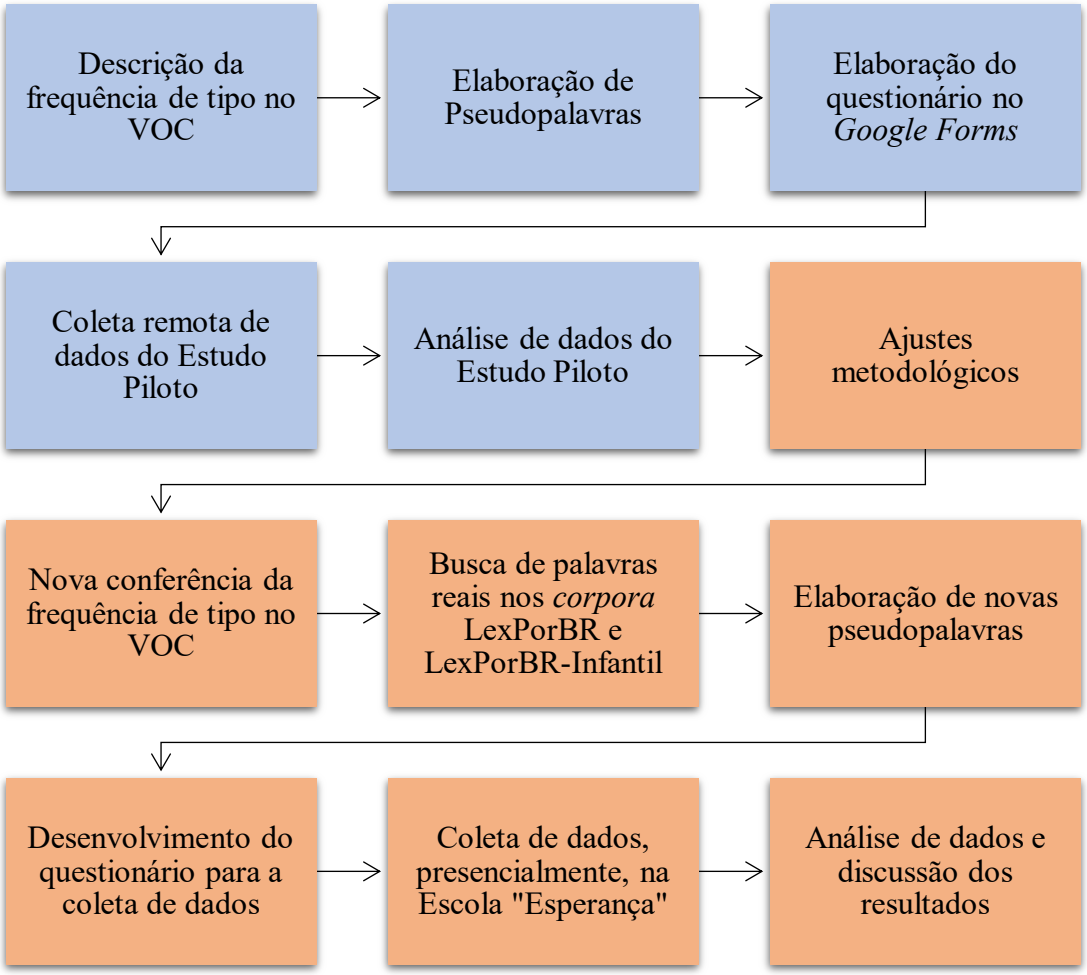
Durante a coleta de dados da "Tarefa 1 – Pseudopalavras", os participantes foram questionados sobre o motivo que os levou a escolher determinado grafema para escrever as palavras que ouviram. No momento da escrita das pseudopalavras, a pesquisadora perguntava: *"o que te levou a escolher determinada letra? O que você pensou sobre?"*. Esta parte do experimento foi gravada e os áudios transcritos para um arquivo do *Word*. O anonimato dos participantes foi mantido e foi utilizado apenas o código para se referir ao sujeito da pesquisa. Por exemplo, código "GV3A" indica um aluno "GV" que está no 3º ano do EF.

A partir dos dados obtidos com as transcrições, uma análise qualitativa foi realizada com intuito de responder ao seguinte questionamento: *"o que os alunos pensam ao ouvir uma não-palavra e, para escrevê-la, ter que optar por grafemas que concorrem para representar um som?*

6.3 Fluxograma da metodologia

Para uma melhor visualização do percurso metodológico desta tese, elaboramos o seguinte Fluxograma. A cor azul refere-se ao passo a passo do Teste Piloto, já a cor laranja refere-se à coleta de dados.

Figura 10 - Fluxograma da metodologia



Fonte: Elaboração própria.

7 ANÁLISE DE DADOS E DISCUSSÃO DOS RESULTADOS

Na ortografia do português brasileiro, há casos de relação direta, ou regulares, entre fonema e grafema, como o fonema /p/ e o grafema <p>, por exemplo, na palavra “pato”. Por outro lado, há palavras que dependem do contexto para regular a escrita de determinada palavra. Por exemplo, o grafema <m>, em posição final de sílaba, é escrito diante de <p> e <b>, como em “pomba”. No entanto, há casos de relações irregulares, em que dois grafemas, em um mesmo contexto, concorrem para representar o mesmo fonema. Por exemplo, os grafemas <c> e <s>, em início de palavras diante de <e> e <i>, concorrem para representar o fonema /s/, como em “cebola” e “seleção”.

Trabalhos apontam que uma das maiores dificuldades dos alunos é em relação às irregularidades ortográficas (Sartori; Mendes; Costa, 2015; Marquardt; Busse, 2015; Nobile; Barrera, 2016; Souza 2019; Castro, 2022). Além disso, erros ortográficos são frequentes em textos de alunos ao longo da educação brasileira e após esta (Sartori; Mendes; Costa, 2015; Selle, 2017; Oliveira *et al.*, 2021; Lacerda; Couto; Oliveira, 2023). Ademais, há indícios de que a frequência, de tipo e de ocorrência, pode atuar neste processo de aprendizagem da escrita ortográfica (Pacton *et al.*, 2001; Santos; Befi-Lopes, 2013; Nigro *et al.*, 2014; Fay; Hein; Ghayoomi, 2015; Ribeiro; Martins, 2020). Neste contexto, esta pesquisa busca defender a tese de que a escrita ortográfica de grafemas concorrentes irregulares é influenciada por múltiplos fatores, como o (1) ano escolar, a (2) relação fonema-grafema, a (3) frequência de tipo e a (4) frequência de ocorrência. Portanto, o objetivo deste Capítulo é analisar os dados coletados e discutir os resultados para a defesa desta hipótese.

Ao longo deste Capítulo, serão analisados os dados coletados na escola pública “Esperança” em quatro momentos da Educação Básica, a saber: 3º ano do Ensino Fundamental I (EFI); 6º ano do Ensino Fundamental II (EFII); 9º ano do Ensino Fundamental II (EFII); 3ª série do Ensino Médio (EM). Três casos de irregularidades ortográficas foram considerados nesta coleta, a saber: o fonema /s/ representado pelos grafemas <c> e <s> em contexto inicial de palavra diante de <e> e <i>; o /ʒ/ representado pelos grafemas <g> e <j> diante de <e> e <i>; e o fonema /z/ em relação a <s> e <z> em contexto intervocálico.

Este Capítulo está organizado em duas partes gerais, a primeira trata-se da análise quantitativa dos dados e a segunda da análise qualitativa. Aquela refere-se à análise dos dados de escrita dos participantes, ou seja, ao que eles escreveram na Tarefa 1 – pseudoplaavras e na Tarefa 2 – palavras reais. Já a segunda refere-se às transcrições das respostas dos alunos em relação ao motivo que os levaram a escrever determinado grafema.

7.1 Análise quantitativa dos dados: da análise à discussão dos resultados

O objetivo desta seção é realizar uma análise quantitativa dos dados, por meio do *software* R Core Team (2023), e discutir os resultados. Esta seção está organizada em duas partes, a saber: "Frequência de tipo: pseudopalavras" e "Frequência de ocorrência: palavras reais". Na primeira, há a análise descritiva e estatística dos dados coletados na "Tarefa 1 – pseudopalavras". Na segunda parte, há a descrição e análise estatística dos dados coletados na "Tarefa 2 – palavras reais.

Como explicado no Capítulo da metodologia, optamos por fazer a análise das palavras reais separadamente das pseudopalavras, pois a forma como a memória é acessada no momento de uma pesquisa linguística pode diferenciar dependendo do estímulo recebido (Cortez, 2018; Bonini; Keske-Soares, 2018). Além disso, optamos, também, pela análise apenas das pseudopalavras, que nos permitiu medir o impacto da frequência de tipo na escrita dos alunos sem a influência da frequência de ocorrência (Cortez, 2018; Bonini; Keske-Soares, 2018). Nas subseções seguintes, as análises descritivas e estatísticas são realizadas.

7.1.1 Frequência de tipo: pseudopalavras

O uso de pseudopalavras, ou não-palavras, é um recurso presente em pesquisas linguísticas, pois, por meio delas, os sujeitos da pesquisa não podem se basear na escrita de palavras reais previamente memorizadas (Altmiller; Treiman; Kessler, 2023). Para observar a atuação da frequência de tipo na escrita de grafemas concorrentes irregulares, foram elaboradas 24 pseudopalavras. Os participantes ouviram, aleatoriamente, cada uma das não-palavras e tiveram que escrevê-las. Como a escrita foi de livre escolha, ou seja, eles escreveram da forma que julgaram correta, todos os dados coletados foram considerados na análise. Nas subseções seguintes, será realizada a análise de dados das pseudopalavras relacionadas a cada um dos fonemas e os grafemas que os representam.

7.1.1.1 Fonema /s/ e os grafemas <c> e <s>

O fonema /s/, articulatoriamente descrito como fricativa sibilante alveolar desvozeada, estabelece uma relação irregular com os grafemas que o representam. Por exemplo, o fonema

/s/ pode ser representado, na escrita, por nove possibilidades de grafemas, como <s>, <c>, <ss>, <c>, <sc>, <sç>, <x>, <xc> e <ç>. Nesta tese, foco é dado ao fonema /s/ em contexto inicial de palavra diante de <e> e <i>, que pode ser representado pelos grafemas <c> e <s>, como nas palavras “situação”, “cilada”, “cebola”, “seguro”. Portanto, para representar o fonema /s/, os grafemas <c> e <s> concorrem quando estão em início de palavra diante de <e> e <i>. A frequência dos grafemas <c> e <s> no contexto descrito foi conferida no Vocabulário Ortográfico Comum da Língua Portuguesa –VOC (Bechara, 2017). Neste contexto inicial de palavra diante de <e> e <i>, o <c> ocorre 3682 vezes e o <s> 5666 vezes. Como apresentado no Capítulo metodológico desta tese, o grafema <s> tem maior frequência de tipo e o <c> menor. Portanto, no contexto inicial de palavras diante de <e> e <i>, grafema <s> ocorre 1,5 a mais do que o <c>. Após a seleção da frequência, as seguintes pseudopalavras foram elaboradas considerando o mesmo contexto. No quadro, a seguir, há as pseudopalavras elaboradas.

Quadro 14 – Pseudopalavras com o fonema /s/

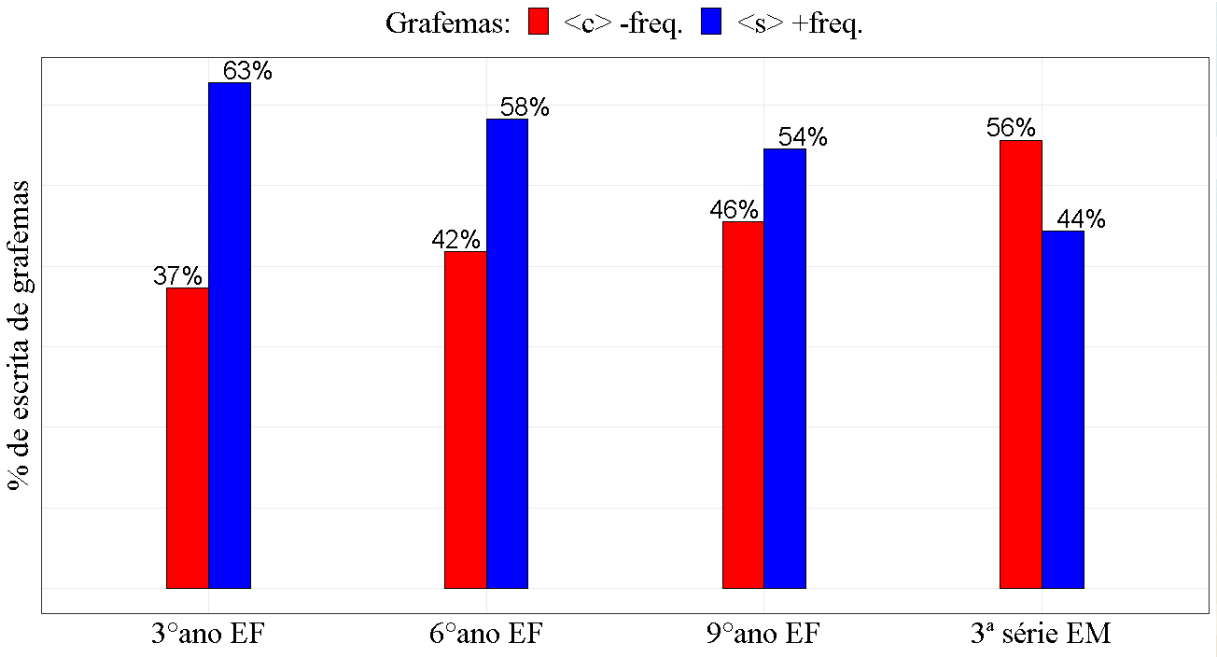
Pesudopalavras	
['sipə]	['sevʊ]
['silʊ]	['sehɪ]
['sibɪ]	['setʊ]
['sikə]	['senɪ]

Fonte: Elaboração própria.

Ao ouvirem tais pseudopalavras, esperávamos que os participantes iniciassem a escrita com o grafema <c> ou o <s>. Nesta primeira parte de análise, consideraremos apenas os dados em que a escrita foi <c> ou <s>. Os dados que fugiram deste esperado serão analisados posteriormente.

No Gráfico 1, a seguir, há, no eixo Y, a porcentagem de escrita dos grafemas e, no eixo X, os anos escolares. A cor azul se refere ao grafema <s> de maior frequência, e a vermelha ao grafema <c> de menor frequência.

Gráfico 1 – Fonema /s/ e a escrita dos grafemas <c> e <s> nas pseudopalavras



Fonte: elaboração própria.

O grafema <c>, menor frequência (barra vermelha), ocorreu em 37% (N = 47) dos dados no 3º ano do EF; em 42% (N = 61) dos dados no 6º ano do EF; em 46% (N = 71) dos dados no 9º do EF; e em 56% (N = 84) dos dados na 3ª série do EM. Já o grafema <s>, maior frequência (barra azul), ocorreu em 63% (N = 79) dos dados no 3º ano do EF; em 58% (N = 85) dos dados no 6º ano do EF; em 54% (N = 85) dos dados no 9º ano do EF; e, por fim, em 44% (N = 67) dos dados na 3ª série do EM. A seguir, há exemplos da escrita realizada pelos participantes, como: “sevo”, “cevo”, “cibi”, “silu”, “selo”, “celo”, “cipa”, “ceni” e “cipa”.

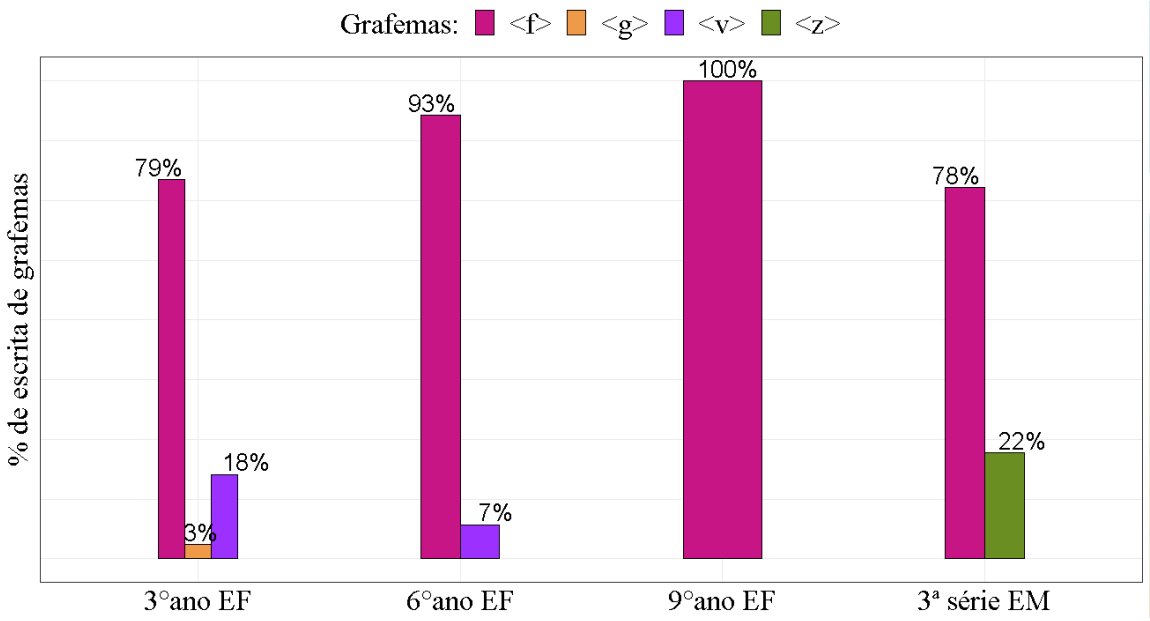
Para observar se há diferença significativa entre a escrita de <s> e <c> em cada ano escolar, realizamos o teste Qui-quadrado comparando os dados do 3º ano do EF e os dados do 6º ano do EF (X^2-squared = 13.235, df = 3, p-value = 0.004). O teste comprovou que há diferença entre o conjunto de dados do 3º ano do EF e do 6º ano do EF. Se há diferença de dados entre dois anos escolares com menor porcentagem de diferença, consequentemente podemos generalizar e afirmar que a ocorrência de escrita dos grafemas é diferente em cada ano escolar.

Nos 3º, 6º e 9º anos, podemos observar que a maior porcentagem de escrita se trata do grafema de maior ocorrência, o <s>. Apenas na 3ª série do EM houve uma tendência à escrita da menor frequência de tipo, o grafema <c>. Investigamos a vogal seguinte ao som inicial do estímulo, o [e] ou [i], pois tínhamos a hipótese de que o contexto poderia influenciar a escolha e de <c> ou <s>. No entanto, não encontramos indícios desta influência.

No que se refere ao fonema /s/ e os grafemas <s> e <c> em contexto inicial de palavras diante de <e> e <i>, podemos argumentar que a escrita de um e outro grafema é diferente em cada ano escolar. Além disso, há uma tendência pela escolha do grafema de maior frequência na maioria dos anos escolares, apenas na 3ª série EM não houve esta inclinação, o que hipoteticamente pode ser explicado pelo fato de os estudantes mais velhos terem mais contato com um léxico maior de palavras e possivelmente terem suas hipóteses iniciais sobre maior frequência colocadas à prova. Este resultado poderá ser encontrado na análise da escrita dos grafemas que representam o fonema /z/, visto que os alunos mais velhos utilizam, no momento da escrita, a relação direta entre som e letra, sem partir da possibilidade de que um grafema pode representar diferentes sons. Portanto, os dados de escrita dos grafemas <s> e <c> confirmam o que outros trabalhos já demonstraram de que a frequência de tipo tem uma atuação na escrita dos alunos (Treiman, 1993; Huback, 2007; Ribeiro; Miranda, 2020; Treiman; Kessler, 2022).

Os dados que fugiram do esperado, ou seja, pseudopalavras em que o participante não escreveu um dos grafemas (<c> ou <s>), foram analisados separadamente na categoria outros. No gráfico a seguir, apresentamos a porcentagem de escrita dos grafemas não esperados.

Gráfico 2 - Fonema /s/ e a escrita de outros grafemas nas pseudopalavras



Fonte: *Corpus* da pesquisa.

O grafema <f> (rosa escuro) ocorreu em 79% (N = 27) dos dados no 3º ano do EF; em 93% (N = 13) dos dados no 6º ano do EF; 100% (N = 4) dos dados no 9º do EF; e em 78% (N = 7) dos dados na 3ª série do EM. O grafema <g> (laranja) ocorreu em 3% (N = 1) dos dados

apenas no 3º ano do EF. O grafema <v> (violeta) ocorreu em 18% (N = 6) dos dados no 3º ano do EF; em 7% (N = 1) dos dados no 6º ano do E. Por fim, o grafema <z> (verde escuro) ocorreu em 22% (N = 2) dos dados. A seguir, há exemplos da escrita realizada pelos participantes, como: “feni”, “fibi”, “ferri”, “zelo”, “vemi”, “gica”, “vica” e “fica”, entre outros.

Como as respostas foram de livre escolha, ou seja, eles não marcaram a alternativa que julgavam correta, mas sim escreviam o estímulo ouvido, optamos por analisar os dados que fugiram do esperado. A maior porcentagem de escrita em todos os anos foi em relação à escrita do grafema <f>. Isso pode ter ocorrido pela familiaridade entre os sons [s] e [f]. Esses sons compartilham características semelhantes, como o modo de articulação, são fricativos, e o desvozeamento, sendo que a única diferença entre eles é o ponto de articulação, o [s] é alveolar e o [f] labiodental.

Vale lembrar que escrita do fonema /s/ por outros grafemas pode ocorrer não só pela semelhança fonética, mas também pela busca de supostas palavras reais, ou com padrões próximos a palavras conhecidas. Pesquisas futuras poderão investigar esta relação e a organização do léxico mental.

7.1.1.2 Fonema /ʒ/ e os grafemas <g> e <j>

O fonema /ʒ/, articulatoriamente descrito como fricativa alveolopalatal vozeada, estabelece uma relação irregular com os grafemas que o representam. Por exemplo, o /ʒ/ pode ser representado, na escrita, por duas possibilidades de grafemas, o <g> e <j> diante de <e> e <i>. Portanto, os grafemas <g> e <j>, quando diante de <e> e <i>, concorrem para representar o fonema /ʒ/. Há casos, por exemplo, “jejum” e “gelo” “jiló” e “girafa”.

Por considerarmos o contexto fonético relevante, nesta tese, o foco é dado ao fonema /ʒ/ diante de <e> e <i> em início de palavras. A frequência dos grafemas <g> e <j> antes de <e> e <i> em início de palavras descrito foi conferida no VOC (Bechara, 2017). Neste contexto inicial de palavra diante de <e> e <i>, o <g> ocorre 1476 vezes e o <j> 444 vezes. Como apresentado no Capítulo metodológico desta tese, o grafema <g> tem maior frequência de tipo e o <j> menor frequência. Portanto, no contexto inicial de palavras diante de <e> e <i>, o grafema <g> ocorre 3,3 vezes a mais do que o <j>. Após a seleção da frequência, as seguintes pseudopalavras foram elaboradas considerando o mesmo contexto.

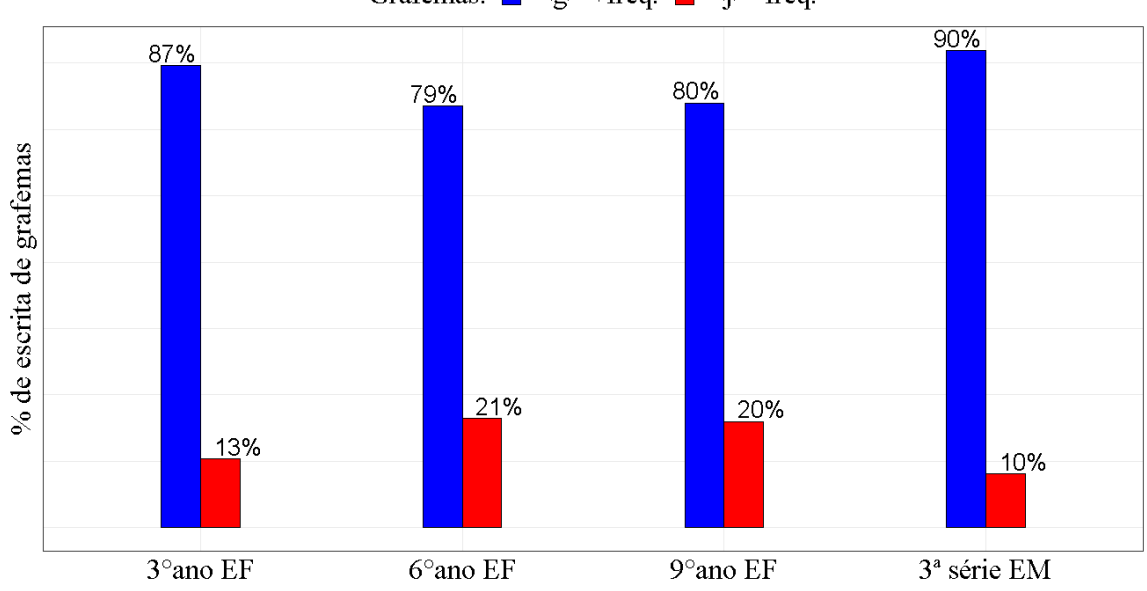
Quadro 15 – Pseudopalavras com o fonema /ʒ/

Pesudopalavras	
['ʒidʃɪ]	['ʒefʊ]
['ʒikʊ]	['ʒetʊ]
['ʒimə]	['ʒepʊ]
['ʒivɪ]	['ʒedʊ]

Fonte: Elaboração própria.

Ao ouvirem tais pseudopalavras, esperávamos que os participantes iniciassem a escrita com o grafema <g> ou <j>. No Gráfico 3, a seguir, há, no eixo Y, a porcentagem dos grafemas escritos e, no eixo X, os anos escolares.

Gráfico 3 - Fonema /ʒ/ e a escrita dos grafemas <g> e <j> nas pseudopalavras



Fonte: Corpus da pesquisa.

O grafema <g>, maior frequência de tipo (barra azul), ocorreu em 87% (N = 134) dos dados no 3º ano do EF; em 79% (N = 120) dos dados no 6º ano do EF; em 80% (N = 128) dos dados no 9º do EF; e em 90% (N = 141) dos dados na 3ª série do EM. Já o grafema <s>, menor frequência (barra vermelha), ocorreu em 13% (N = 20) dos dados no 3º ano do EF; em 21% (N = 31) dos dados no 6º ano do EF; em 20% (N = 32) dos dados no 9º ano do EF; e, por fim, em 10% (N = 16) dos dados na 3ª série do EM. A seguir, há exemplos da escrita realizada pelos participantes, como: “gico”, “gima”, “givi”, “jefo”, “jidi”, “jico”, “gidi”, “giko” e “jepo”.

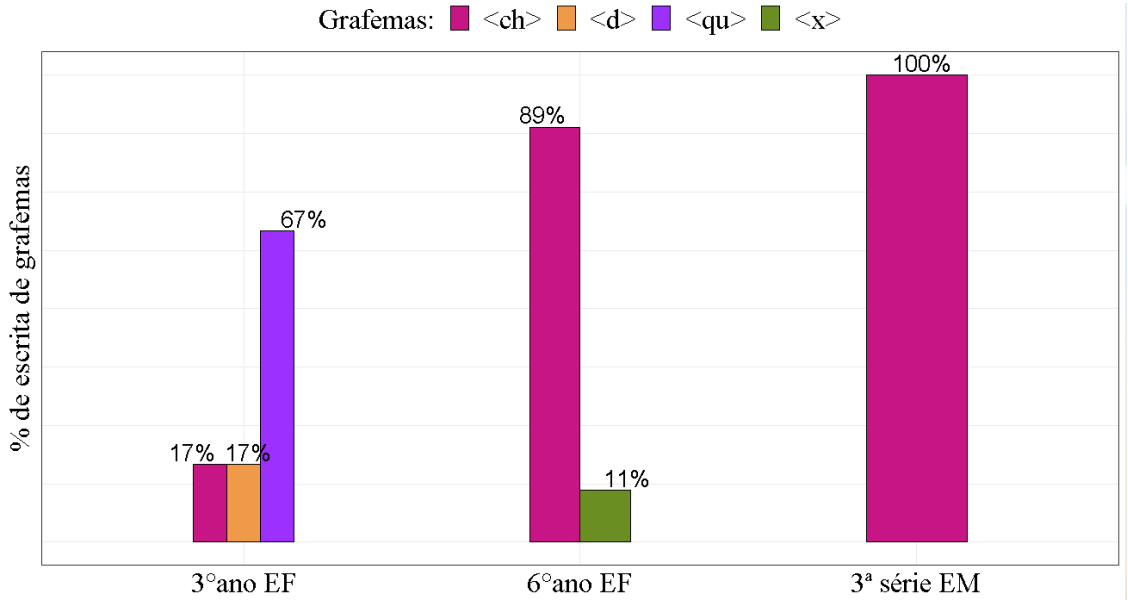
Para afirmar se há diferença significativa entre a escrita de <g> e <j> em cada ano escolar, realizamos o teste Qui-quadrado comparando os dados do 6º ano do EF e os dados do

9º ano do EF (X^2-squared = 110.47, df = 3, p-value < 2.2e-16). O teste comprovou que há diferença entre o conjunto de dados do 6º ano do EF e do 9º ano do EF. Se há diferença de dados entre dois anos escolares com menor porcentagem de diferença, consequentemente podemos generalizar e afirmar que a ocorrência de escrita dos grafemas é diferente em cada ano escolar.

No português brasileiro, o grafema <g> ocorre 3,3 vezes a mais do que o <j>. Em consequência disso, ao observar os dados em cada ano escolar, podemos afirmar que os alunos têm preferência por escrever <g>, de maior frequência de tipo, do que o <j>. Os dados de escrita dos grafemas <g> e <j> confirmam o que outros trabalhos já demonstraram de que a frequência de tipo tem uma atuação na escrita dos alunos (Treiman; Kessler, 2014; Treiman *et al*, 2018; Huback, 2007; Guimarães; Cristófaro Silva; Gomes, 2020; Ribeiro; Martins, 2020). Portanto, podemos afirmar que, na concorrência entre os grafemas <g> e <j> para representar o fonema /ʒ/, há uma tendência à escolha do mais frequente, o que confirma também a atuação do conhecimento estatístico (Treiman, 1993, 2017; Treiman; Kessler, 2014; Chetail, 2017; Treiman *et al*. 2019; Treiman; Kessler, 2022).

Os dados que não seguiram o padrão esperado, ou seja, dados em que o participante não escreveu um dos grafemas esperados (<g> ou <j>), foram analisados separadamente na categoria outros. No Gráfico 4, a seguir, há a porcentagem de ocorrência dos grafemas não esperados.

Gráfico 4 - Fonema /ʒ/ e a escrita de outros grafemas nas pseudopalavras



Fonte: *Corpus* da pesquisa.

O grafema <ch> (rosa escuro) ocorreu em 17% (N = 1) dos dados no 3º ano do EF; em 93% (N = 13) dos dados no 6º ano do EF; e em 100% (N = 3) dos dados na 3ª série do EM. O grafema <d> (laranja) ocorreu em 17% (N = 1) dos dados apenas no 3º ano do EF. O grafema <qu> (violeta) ocorreu em 67% (N = 4) dos dados apenas no 3º ano do EF. Por fim, o grafema <x> (verde escuro) ocorreu em 11% (N = 1) dos dados apenas no 6º ano EF. No 9º ano, não houve nenhuma ocorrência de escrita de grafemas não esperados. A seguir, há exemplos da escrita realizada pelos participantes, como: “chivi”, “chefu”, “xerfu” e “didi”, “quidi”, entre outros.

Nos estímulos do fonema /ʒ/, podemos observar que foram poucos casos de escrita de grafemas não esperados. A maior porcentagem de escrita em todos os anos escolares foi em relação à escrita do dígrafo <ch>. Isso pode ter ocorrido pela familiaridade entre os sons [ʒ] e [ʃ], que pode ser representado por <ch> e <x>. Esses sons compartilham características semelhantes, como o modo de articulação (fricativas) e o ponto de articulação (alveolopalatais) são fricativas. A única diferença entre eles é o vozeamento, o [ʒ] é vozeado e o [ʃ] desvozeado., o que indica que a percepção fonética ainda está se consolidando para estes estudantes. Para a escrita dos grafemas <d> e <qu>, não há motivação aparente, além da possível busca por um padrão familiar, tanto foneticamente, quanto semanticamente.

7.1.1.3 Fonema /z/ e os grafemas <z> e <s>

O fonema /z/, articulatoriamente descrito como fricativa sibilante alveolar vozeada, estabelece uma relação irregular com os grafemas que o representam. Por exemplo, o /z/ pode ser representado, na escrita, por três possibilidades de grafemas, o <s>, o <x> e o <z>. Portanto, os grafemas <s>, <z> e <x> concorrem para representar o fonema /z/, por exemplo, “asilo” e “azul” e “exame”.

Existe um número baixo, no português, de palavras escritas com <x> representando o som [z]. Por outro lado, há mais casos dos grafemas <s> e <z>, entre vogais, que concorrem para representar o fonema /z/, do que o grafema <x>. Por considerarmos o contexto importante na linguística, nesta tese, o foco é dado aos grafemas <s> e <z> entre vogais que representam o fonema /z/. A frequência dos grafemas <s> e <z> no contexto descrito foi conferida no VOC (Bechara, 2017). Neste contexto entre vogais, o <s> ocorre 110597 vezes e o <z> 92033 vezes. Como apresentado no Capítulo metodológico desta tese, o grafema <s> tem maior frequência de tipo e o <z> menor frequência. Portanto, no contexto intervocálico, grafema <s> ocorre 1,2

vezes a mais do que o <z>. Após a seleção da frequência, as seguintes pseudopalavras foram elaboradas considerando o mesmo contexto.

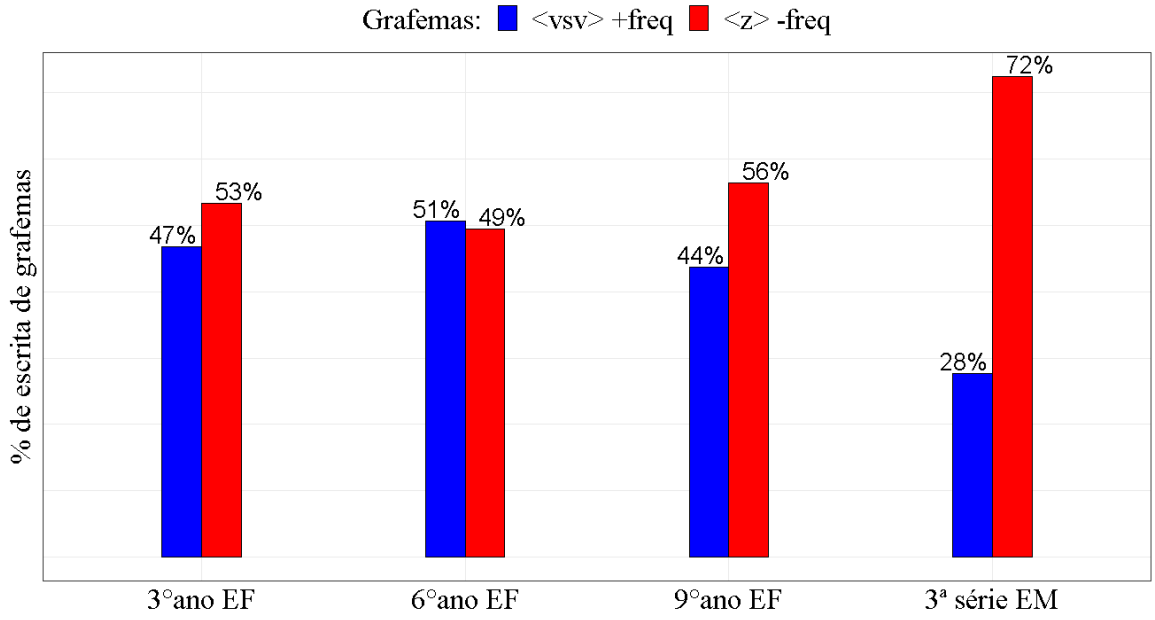
Quadro 16 – Pseudopalavras com o fonema /z/

Pesudopalavras	
['tezə]	['dʒizə]
['bezə]	['hizə]
['kezə]	['kizə]
['fezə]	['nizə]

Fonte: Elaboração própria.

Ao ouvirem tais pseudopalavras, esperávamos que os participantes iniciassem a escrita com o grafema <VsV> ou <z>. Para diferenciar o grafema <s> que representa o fonema /s/, analisado nas seções anteriores, optamos por utilizar <VsV>, o grafema <s> entre vogais que representa o fonema /z/. No Gráfico 5, a seguir, há, no eixo Y, a porcentagem dos grafemas escritos e, no eixo X, os anos escolares.

Gráfico 5 - Fonema /z/ e a escrita dos grafemas <VsV> e <z> nas pseudopalavras



Fonte: Elaboração própria.

O grafema <VsV>, maior frequência de tipo (barra azul), ocorreu em 47% (N = 71) dos dados no 3º ano do EF; em 51% (N = 81) dos dados no 6º ano do EF; em 44% (N = 69) dos dados no 9º do EF; e em 28% (N = 44) dos dados na 3ª série do EM. Já o grafema <z>, menor frequência (barra vermelha), ocorreu em 53% (N = 81) dos dados no 3º ano do EF; em 49% (N

= 79) dos dados no 6º ano do EF; em 56% (N = 89) dos dados no 9º ano do EF; e, por fim, em 72% (N = 115) dos dados na 3ª série do EM. A seguir, há exemplos da escrita realizada: “deza”, “tesa”, “beza”, “risa”, “kiza”, “quiza”, “niza”, “fesa” e “tesa”.

Para afirmar se há diferença significativa entre a escrita de <VsV> e <z> em cada ano escolar, realizamos o teste Qui-quadrado comparando os dados do 9º ano do EF e os dados da 3ª série do EF (X^2-squared = 12.87, df = 3, p-value =0.004926). O teste comprovou que há diferença entre o conjunto de dados do 9º ano do EF e da 3ª série do EM. Apenas o conjunto de dados da 3ª série que se difere dos demais conjuntos de dados dos 3º, 6º e 9º anos do EF. Não houve diferença estatística nos seguintes conjuntos de dados: 3º e 6º anos (X^2-squared = 0.87179, df = 3, p-value = 0.8322); 6º e 9º anos (X^2 - squared = 2.5535, df = 3, p-value = 0.4657).

Ao analisar o Gráfico 5, podemos observar que há uma preferência pela escrita de <z> (menor frequência) no 3º ano, 9º ano e 3ª série do Ensino Médio. Este resultado contraria um dos fatores da hipótese desta tese de que a frequência de tipo pode influenciar na escrita de grafemas concorrentes irregulares. Neste caso, o aluno tenderia a escrever o grafema de maior frequência, e não o de menor, como os resultados desta seção sugerem. Este dado é evidência de que a seleção ortográfica em caso de palavras inventadas pode não perpassar pela frequência ou mesmo relacionar o conhecimento estatístico a padrões não avaliados nesta tese. Ainda pesquisas futuras podem investigar a interação entre frequência com outros parâmetros, por exemplo, o feixe de possibilidades escrita que cada grafema dispõe.

Nas análises qualitativas dos dados, que serão realizadas ao final deste Capítulo, há indícios de que os participantes da 3ª série do EM partiram da relação direta entre som e grafema. Ou seja, ao ouvirem o estímulo com [z], automaticamente escreveram a letra <z>. Nas justificativas dessa escrita, eles argumentaram que não escreveram a letra <s>, pois representaria o som [s], e não o [z]. Portanto, podemos inferir que, antes da frequência, os alunos utilizaram a relação direta entre fonema e grafema para a escrita das pseudopalavras. Isso vai ao encontro dos pressupostos da IMP ao afirmar que a aprendizagem da escrita ortográfica é múltipla, que pode haver a atuação de diversos padrões, como os fonológicos, gráficos, entre outros (Treiman; Kessler, 2014; Treiman *et al*, 2018; Castro; Couto, 2021; Oliveira; Castro; Couto, 2023).

Os dados que não seguiram o esperado, ou seja, casos em que o participante não escreveu um dos grafemas esperados (<s> ou <z>) forma poucos, no total de 11 ocorrências. Por esse baixo número, optamos por não os discutir. Na seção seguinte, uma discussão geral dos dados relacionados à frequência de tipo é realizada.

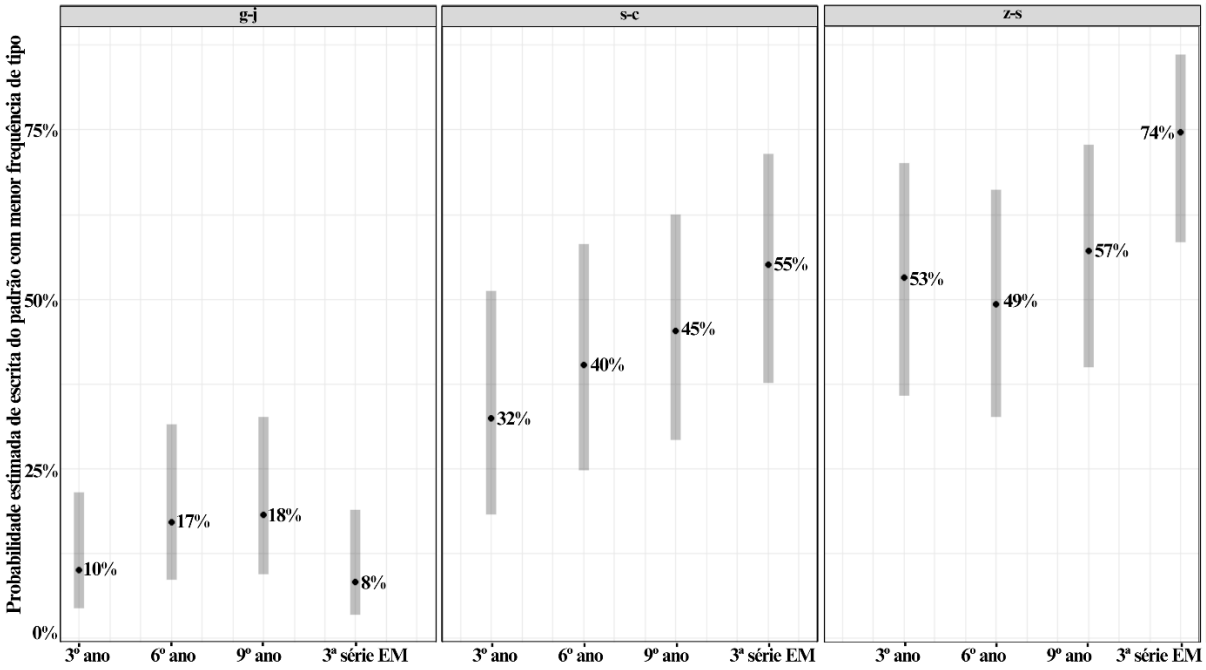
7.1.1.4 Discussão geral dos resultados da frequência de tipo - pseudopalavras

Esta pesquisa busca defender a tese de que a escrita ortográfica de grafemas concorrentes irregulares é influenciada por múltiplos fatores, como o (1) ano escolar, a (2) relação fonema-grafema, a (3) frequência de tipo e a (4) frequência de ocorrência. Nesta seção, investigamos se o ano escolar, a relação fonema-grafema e a frequência de tipo influenciam na escrita de grafemas concorrentes irregulares. Para isso, um modelo linear generalizado misto[22] foi ajustado à amostra de dados das pseudopalavras. A função do modelo foi de estimar a probabilidade de ocorrência de grafemas de menor frequência de tipo por ano escolar.

A variável dependente do modelo foi categórica: frequência de tipo (maior/menor). As variáveis independentes de efeito fixo do modelo foram o ano escolar (3A, 6ª, 9ª e 3EM) e o grafema (s-c, g-j e z-s). Além disso, o participante foi incluído ao modelo como variável independente de efeito aleatório. Uma comparação de modelos aninhados mostrou que o melhor modelo ajustado considera o grafema em interação com o ano escolar ($X^2 = 31.804$, p-value = 0.0002153). Em outras palavras, o melhor modelo ajustado aos dados indica que a escrita ortográfica do grafema de menor frequência de tipo é influenciada pelo ano escolar em interação com a relação fonema-grafema. No Gráfico 6, há a probabilidade estimada pelo modelo em relação à escrita de grafemas concorrentes irregulares por ano escolar.

[22] mmpgeral.grafemaano.interacao <-glmer(ftipoescritacat~grafema*ano+(1|participante), dados.pseudo, family = binomial)

Gráfico 6 - Probabilidade de escrita do grafema com a menor frequência de tipo



Fonte: Elaboração própria.

O Gráfico 6 está organizado da seguinte forma: a probabilidade de escrita dos grafemas de menor frequência de tipo está no eixo Y. No eixo X, há os anos escolares. O retângulo à esquerda nos mostra a probabilidade de ocorrência do grafema <j> (menor frequência) em cada ano escolar. O retângulo do meio, por sua vez, nos dá a probabilidade de ocorrência do grafema <c> (menor frequência) por ano escolar. O retângulo à direita mostra a probabilidade de ocorrência do <z> (menor frequência) também em cada ano escolar analisado. É importante lembrar que os grafemas <g> e <j>, diante de <e> e <i>, concorrem para representar o fonema /ʒ/; os grafemas <c> e <s>, em início de palavra e diante de <e> e <i>, competem para representar o fonema /s/; e os grafemas <z> e <VsV> em contexto intervocálico concorrem para representar o fonema /z/.

Podemos observar que a escrita do grafema <j> é a menos provável em todos os anos escolares. Por exemplo, a probabilidade de os alunos do 3° ano do EF escreverem o grafema <j> em vez de <g> é de apenas 10%. Em relação à escrita do grafema <c>, em vez de <s>, por outro lado, a probabilidade aumenta para 34% nos dados do 3° ano do EF. No que diz respeito aos grafemas <z> e <s>, há 53% de probabilidade de ocorrer <z>, em vez de <VsV>, nos dados 3° ano do EF. Em suma, a escrita do grafema parece depender não apenas da baixa frequência de tipo, como também da relação fonema-grafema.

Vale ressaltar ainda que, além da probabilidade estimada pelo modelo, o Gráfico 6 nos mostra o intervalo de confiança. Segundo Oushiro (2017), o intervalo de confiança "estabelece um valor mínimo e um valor máximo em que se calcula estar o verdadeiro parâmetro da população". Em outras palavras, o intervalo de confiança calcula as incertezas relacionadas à probabilidade estimada do modelo. Quanto menor o intervalo de confiança, maior a previsibilidade do modelo. Ao observarmos a sombra cinza no Gráfico 6, podemos perceber que, mesmo considerando os intervalos de confiança, as menores probabilidades se relacionam à ocorrência do <j>. Ora, por que a probabilidade de ocorrência do grafema <z> é menor do que a dos grafemas <c> e <z>, se todos esses grafemas têm baixa frequência de tipo? Esperávamos que a probabilidade de ocorrência do grafema com menor frequência de tipo seria baixa nos três conjuntos.

Uma possível resposta a essa pergunta pode estar relacionada ao fato de que as taxas das frequências de tipo variam a depender do grafema. Por opção metodológica, categorizamos os grafemas <j>, <s> e <z> como "grafemas de baixa frequência". Contudo, é necessário pontuar que, cada grafema tem uma frequência de tipo específica, como <j> (444), <c> (3682) e <z> (92033). A frequência de tipo do grafema <j> é a menor analisada nesta pesquisa. Consequentemente, é perceptível que o grafema <j> é o menos provável de ocorrer nos dados do 3°, 6° e 9° anos do Ensino Fundamental, e na 3ª série do Ensino Médio.

Podemos afirmar, portanto, que a escrita do grafema com menor frequência de tipo varia a depender de outros fatores, com a relação fonema-grafema. Ou seja, não podemos considerar a frequência de tipo isoladamente, é importante levar em consideração a relação fonema-grafema e o ano escolar.

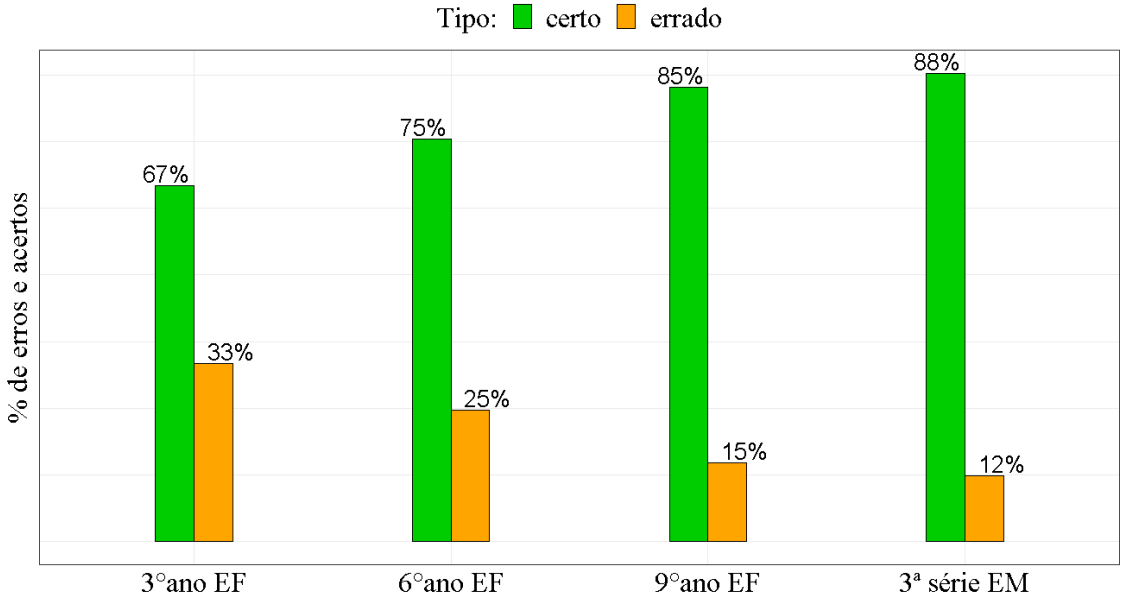
7.1.2 Frequência de ocorrência: palavras reais

Nesta seção, analisamos os dados relativos às palavras reais coletados na "Tarefa 2 – Palavras reais". Assim como nos estímulos da Tarefa 1, nesta os participantes ouviram, aleatoriamente, e escreveram as palavras escutadas. As palavras deste experimento foram selecionadas em dois *corpora*, o LexPorBR e o LexPorBR – Infantil e foram classificadas quanto à frequência de ocorrência, ou seja, quanto ao número de vezes com que elas ocorrem nos dois *corpora*. Como são palavras que existem, trabalharemos com os termos certo e errado de acordo com as regras ortográficas do português brasileiro. Além disso, consideramos certo ou errado apenas no que diz respeito aos grafemas analisados. Por exemplo, se o participante

escreveu *jeep no lugar de “jipe”, consideramos como certo, pois o estudante, apesar de ter errado na escrita, ele acertou em relação ao grafema <j>.

No Gráfico 7 a seguir, há a porcentagem de erros e acertos em cada ano escolar.

Gráfico 7 – Erros e acertos na escrita das palavras reais



Fonte: *Corpus* da pesquisa.

No eixo X do gráfico 7, há os anos escolares e no Y a porcentagem de ocorrência de acertos (barras verdes) e erros (barras laranjas). O gráfico 7 nos mostra que, no 3º ano do EF, 67% (N = 320) dos dados são acertos e 33% (N = 160) de erros. No 6º ano do EF, há 75% (N = 362) de acertos e 25% (N = 118) de erros. No 9º ano, por sua vez, há 85% (N = 409) de acertos e 15% (N = 71) de erros. Por fim, na 3ª série do EM, há 88% (N = 421) de acertos e 12% (N = 59) de erros. A título de exemplificação, apresentamos os seguintes erros: *serne (“cerne”), “cingelo (“singelo”), *gipe (“jipe”), *jestual (“gestual”), *turqueza (“turquesa”), *amisade (“amizade”), entre outros.

Para afirmar se há diferença significativa entre erros e acertos em cada ano escolar, realizamos o teste Qui-quadrado comparando os dados do 9º ano do EF e os dados da 3ª série do EF (X^2- squared = 511.02, df = 3, p-value < 2.2e-16). O teste comprovou que há diferença entre o conjunto de dados do 9º ano do EF e da 3ª série do EM. Se há diferença de dados entre dois anos escolares com menor porcentagem de diferença, consequentemente podemos generalizar e afirmar que cada conjunto de dados é diferente em cada ano escolar. É interessante observar que de um ano escolar para o outro, há o aumento de acertos e a redução de erros

ortográficos. Isso pode ser uma evidência do impacto positivo da escola e do maior contato com letramento na escrita ortográfica dos alunos.

Vale ressaltar que, mesmo que percebamos uma redução nos erros ortográficos em cada série escolar, ainda há a ocorrência deles. Diante disso, optamos por investigar tais erros, com intuito de analisar se a frequência de tipo e de ocorrência podem ter relação com os erros ortográficos. Nas subseções seguintes, analisaremos os dados de escrita das palavras reais por fonemas e os grafemas que o representa.

7.1.2.1 Fonema /s/ e os grafemas <c> e <s>

Os grafemas <c> e <s> em início de palavras e diante de <e> e <i> concorrem para representar o fonema /s/. Há, por exemplo, as palavras “situação”, “cilada”, “cebola”, “seguro”. Por decisão metodológica, classificamos, nesta tese, o grafema <c> de menor frequência e o <s> de maior frequência, ambos no contexto descrito. Para a Tarefa 2 – palavras reais, selecionamos quatro palavras iniciadas com <c> e quatro iniciadas com <s> no LexPorBR e LexPorBR – Infantil. Duas dessas quatros palavras possuem maior frequência de ocorrência e as outras duas, menor frequência de ocorrência. O Quadro 17 apresenta essas palavras, que estão organizadas em maior e menor frequência de ocorrência e em relação aos grafemas <s> e <c>.

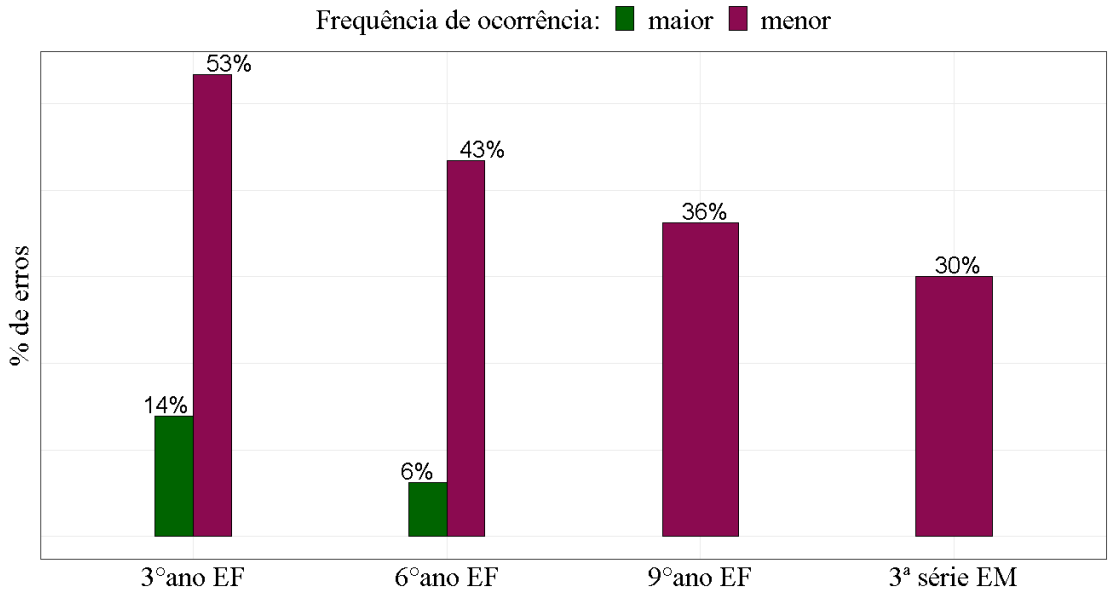
Quadro 17 – Palavras reais com os grafemas <c> e <s> que representam o fonema /s/

Frequência de ocorrência	**Palavras**			
	Grafema <c> (menor frequência de tipo)		**Grafema <s>** (maior frequência de tipo)	
Maior	cidade	certo	semana	sistema
Menor	cicuta	cerne	sequela	singelo

Fonte: Elaboração própria.

Ao ouvirem os estímulos das palavras reais e observarem as imagens que remetiam a elas, os sujeitos da pesquisa escreviam de acordo com o que eles julgaram ser a forma correta. Em relação a estes dados, filtramos e analisamos apenas os erros ortográficos. No Gráfico 8, a seguir, há a porcentagem de erros ortográficos por ano escolar e pela frequência de ocorrência.

Gráfico 8 – Fonema /s/ e os erros na escrita das palavras reais com <c> e <s>



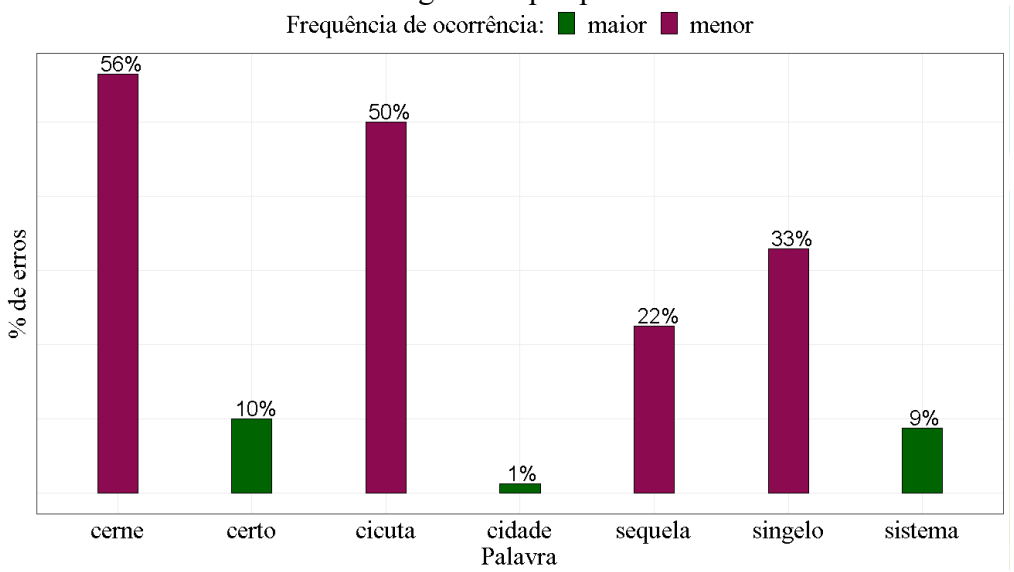
Fonte: *Corpus* da pesquisa.

No eixo X, há os anos escolares e o no eixo Y a porcentagem de erros ortográficos. As barras de cor roxa são referentes aos erros ortográficos em palavras de menor frequência de ocorrência, já as barras de cor verde são erros em palavras de maior frequência de ocorrência. Nas palavras de maior frequência (cor verde escuro), ocorreram erros ortográficos em 14% (N = 11) dos dados no 3º ano do EF; e em 6% (N = 5) no 6º ano do EF. Já nas palavras de menor de ocorrência frequência (cor roxa), foram encontrados erros ortográficos em 53% (N = 40) dos dados no 3º ano do EF; em 43% (N = 33) dos dados no 6º ano; em 36% (N = 29) dos dados no 9º ano; e em 30% (N = 24) dos dados na 3ª série do EM.

Ao analisar o Gráfico 8, é perceptível que os erros relacionados à irregularidade ortográfica do <c> e <s> ocorrem mais em palavras de baixa frequência em cada ano escolar. Aplicamos um teste Qui-quadrado que comprovou que há diferença entre o índice de erros ortográficos em palavras de menor e maior frequência de ocorrência (X^2-squared = 97.549, df = 7, p-value < 2.2e-16). Os alunos, portanto, são sensíveis à frequência de palavras e resolvem os problemas de escrita em palavras de maior frequência de ocorrência. Isso corrobora pesquisas realizadas que concluíram que os alunos, em idade escolar, erram mais em palavras infrequentes (Santos; Befi-Lopes, 2013; Ribeiro; Martins, 2020). Portanto, não podemos generalizar que os alunos têm mais dificuldades em relação às palavras de ortografia irregular, pois, ao olhar mais de perto os erros ortográficos, observamos que erros ocorrem em palavras de ortografia irregular, mas de baixa frequência. Ou seja, há interação entre dois fatores: frequência e relações ortográficas.

Além disso, é importante investigar em quais palavras ocorrem os erros ortográficos. Para isso, analisaremos o Gráfico 9, em que há a porcentagem de erros ortográficos no eixo Y, e por palavras no eixo X.

Gráfico 9 – Erros ortográficos por palavras com <c> e <s>



Fonte: Elaboração própria.

As barras de cor roxa são referentes aos erros ortográficos em palavras de menor frequência de ocorrência, já as barras de cor verde são erros em palavras de maior frequência de ocorrência. Nas palavras de menor frequência de ocorrência (barras roxas), ocorreram 56% (N= 44) erros ortográficos no vocábulo “cerne”; 50% (N= 40) de erros na palavra “cicuta”; 22% (N=18) de erros no vocábulo “sequela”; e 33% (N=24) de erros ortográficos na palavra “singelo”. Já nas palavras de maior frequência (barras verdes), ocorreram 10% (N=8) de erros na palavra “certo”; 1% (N=1) de erro no vocábulo “cidade”; por fim, 9% (N = 7) de erros no vocábulo “sistema”. Na palavra “semana”, não ocorreu erro em nenhum ano escolar, por isso ela não aparece no Gráfico 9.

Aplicamos um teste Qui- quadrado para avaliar a diferença de erros ortográficos nas palavras de maior e menor frequência de ocorrência (X^2 -squared = 114.34, df = 7, p-value < 2.2e-16). O teste confirmou que há diferença entre as palavras de menor e maior frequência de ocorrência. Portanto, podemos afirmar que as maiores porcentagens de erros ortográficos ocorrem em palavras de baixa frequência.

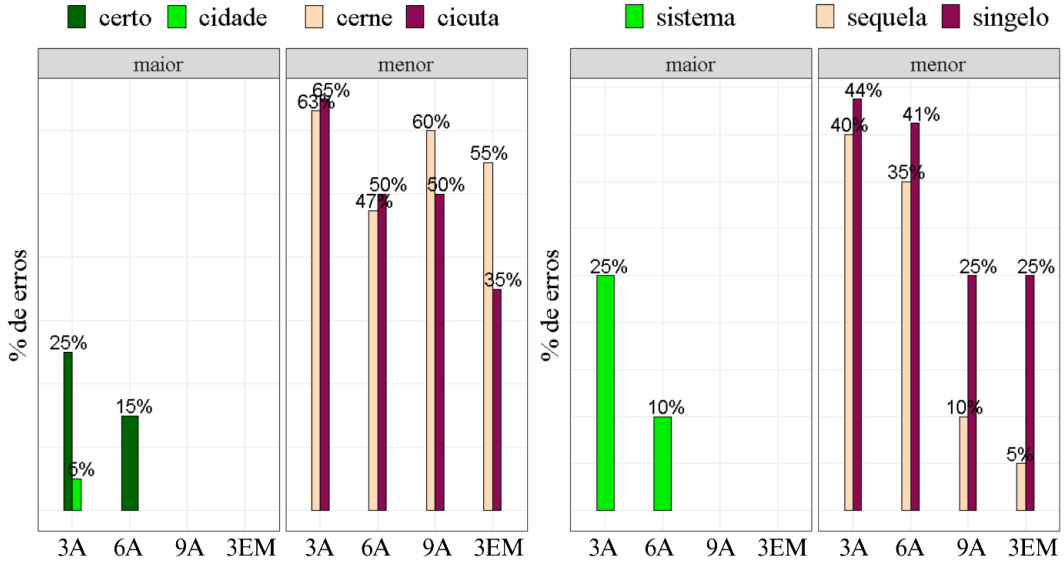
Ao observar o Gráfico 9, especificamente nas palavras de baixa frequência de ocorrência, percebemos que a frequência de tipo parece ter uma atuação em relação aos erros ortográficos, pois a maior porcentagem de erros ortográficos é nas palavras “cerne” e “cicuta”,

com o grafema <c>, que categorizamos como menor frequência de tipo. Aplicamos o teste Qui-quadrado para avaliar a diferença entre os erros ortográficos nas palavras “cerne” e “cicuta” e “sequela” e “singelo” (X^2-squared = 10.769, df = 3, p-value = 0.01304). O teste demonstrou que os índices de erros ortográficos são diferentes nas palavras escritas ortograficamente com <c> (menor frequência de tipo) e com <s> (maior frequência de tipo).

Além disso, verificamos que os participantes quando erram as palavras, que são escritas ortograficamente com <c>, tendem a escreverem com <s> (maior frequência), como em *sicuta (“cicuta”), *serne (“cerne). Portanto, no contexto de menor frequência de tipo (<c>), o usuário da língua tende a recorrer ao padrão da maior frequência de ocorrência (<s>) para grafar palavras de baixa frequência de ocorrência. Isto indica que há uma interação entre as frequências, de tipo e de ocorrência, o que corrobora os resultados de Huback (2007).

No gráfico 10, observaremos a relação entre frequência de tipo e de ocorrência por ano escolar. Podemos ler o Gráfico 10 da seguinte forma: no eixo X, há os anos escolares 3A (3º ano do EF), 6A (6º ano do EF), 9A (9º ano do EF), 3EM (3ª série do EF) e o eixo Y a porcentagem de erros ortográficos. Nos dois retângulos à esquerda, há o grupo de palavras escritas ortograficamente com <c> (menor frequência de tipo). Nos retângulos à direita, há o grupo de palavras escritas ortograficamente com <s> (maior frequência de tipo). Cada grupo de palavras está organizado em relação à frequência de ocorrência. As palavras nas cores verdes (verde escuro e claro) são de maior frequência de ocorrência; nas cores roxo e rosa claro, há as palavras de baixa frequência de ocorrência.

Gráfico 10 - Frequência de tipo de <c> e <s> e frequência de ocorrência por ano escolar e por palavras com erros ortográficos



Fonte: Elaboração própria.

No Gráfico 10, podemos observar que erros ortográficos ocorreram em todas as palavras escritas ortograficamente com <c>. Já nas palavras escrita com <s>, ocorreram erros em três (“sistema”, “sequela” e singelo”) das quatro palavras do experimento. Na palavra “semana”, não ocorreu erro em nenhum ano escolar. Além disso, é nítido observar que os maiores índices de erros ortográficos ocorreram em palavras de menor frequência de ocorrência (cores vinho e rosa claro). Portanto, podemos inferir que a frequência de ocorrência pode ter uma maior atuação do que a de tipo durante o processo de escrita ortográfica.

7.1.2.2 Fonema /ʒ/ e os grafemas <g> e <j>

Os grafemas <g> e <j>, diante de <e> e <i>, concorrem para representar o fonema /ʒ/. Há, por exemplo, as palavras “general” e “jipe”, gigante” e “jiló”. Considerando o padrão comparativo, classificamos, nesta tese, o grafema <j> de menor frequência e o <g> de maior frequência, ambos no contexto descrito. Para a Tarefa 2 – palavras reais, selecionamos quatro palavras iniciadas com <g> e quatro iniciadas com <j> no LexPorBR e LexPorBR – Infantil. Duas dessas quatros palavras possuem maior frequência de ocorrência e as outras duas, menor frequência de ocorrência. O Quadro 18 apresenta essas palavras, que estão organizadas em maior e menor frequência de ocorrência e em relação aos grafemas <g> e <j> no contexto descrito.

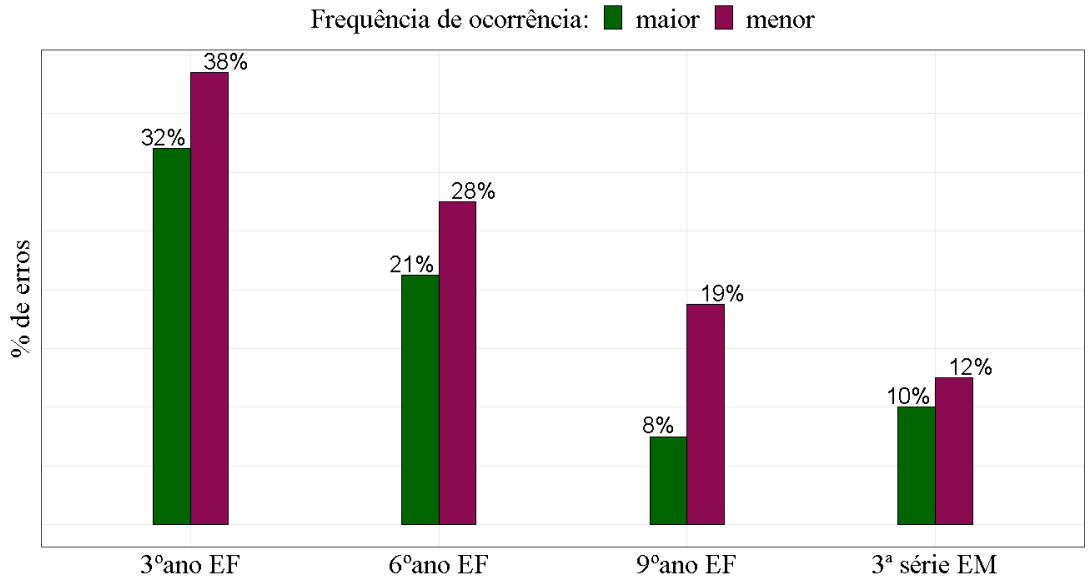
Quadro 18 – Palavras reais com os grafemas <g> e <j> que representam o fonema /ʒ/

Frequência de ocorrência	**Palavras**			
	Grafema <g>		**Grafema <j>**	
Maior	general	gigante	sujeito	jipe
Menor	gestual	gincana	dejeto	jiló

Fonte: Elaboração própria.

Ao ouvirem os estímulos das palavras reais e observarem as imagens que remetiam a elas, os participantes escreviam de acordo com o que eles julgassem ser a forma correta. Em relação a estes dados, filtramos apenas os erros ortográficos e os analisaremos. No Gráfico 11, a seguir, há a porcentagem de erros ortográficos por ano escolar e pela frequência de ocorrência.

Gráfico 11 – Fonema /ʒ/ e os erros na escrita das palavras reais com <g> e <j>

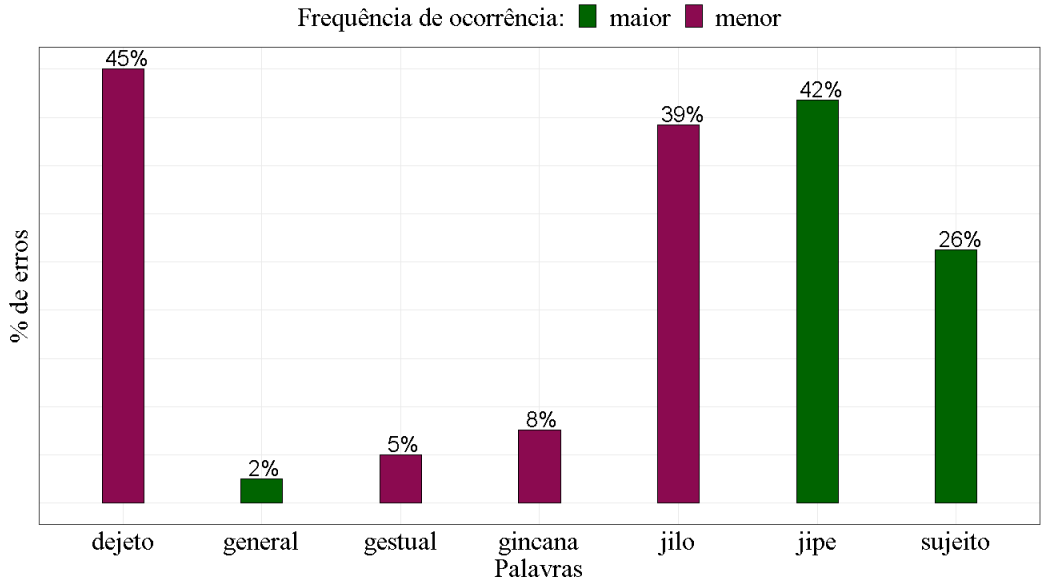


Fonte: *Corpus* da pesquisa.

No eixo X, há os anos escolares e, no eixo Y, a porcentagem de erros ortográficos. As barras de cor roxa são referentes aos erros ortográficos em palavras de menor frequência de ocorrência, já as barras de cor verde são erros em palavras de maior frequência de ocorrência. Nas palavras de maior frequência (cor verde escuro), os erros ortográficos ocorreram em 32% (N = 25) dos dados no 3º ano do EF; e em 21% (N = 17) no 6º ano do EF; em 8% (N = 6) dos dados no 9º ano do EF; e em 10% (N = 8) dos dados na 3ª série do EM. Já nas palavras de menor frequência de ocorrência (cor roxa), foram encontrados erros ortográficos em 38% (N = 30) dos dados no 3º ano do EF; em21% (N = 22) dos dados no 6º ano; em8% (N = 15) dos dados no 9º ano; e 12% (N = 10) dos dados na 3ª série do Ensino Médio.

Ao analisar o Gráfico 11, podemos observar que o percentual de erros ortográficos reduz em cada ano escolar. Além disso, é perceptível que as altas porcentagens de erros ortográficos ocorrem em palavras de menor frequência de ocorrência. Para confirmar se há diferença entre o índice de erros ortográficos em palavras com maior (cor verde escuro) e com menor frequência de ocorrência (cor roxa) em cada ano escolar, aplicamos um teste Qui-quadrado que mostrou que não há diferença significativa entre estes dois grupos de palavras (X^2-squared = 30.789, df = 7, p-value = 6.799e-5). Portanto, diferentemente dos dados do fonema /s/, analisados anteriormente, não há diferença entre os erros ortográficos em palavras de maior ou menor frequência de ocorrência escritas com <g> e <j>. Para compreender melhor este resultado, analisamos em quais palavras ocorreram os erros ortográficos. No Gráfico 12, em que há a porcentagem de erros ortográficos, no eixo Y e no eixo X por palavras.

Gráfico 12 – Erros ortográficos por palavras com <g> e <j>



Fonte: Elaboração própria.

As barras de cor roxa são referentes aos erros ortográficos em palavras de menor frequência de ocorrência, já as barras de cor verde são erros em palavras de maior frequência de ocorrência. Nas palavras de menor frequência de ocorrência (barras roxas), ocorreram 45% (N= 36) erros ortográficos no vocábulo “dejeto”; 5% (N= 2) de erros na palavra “gestual”; 8% (N=6) de erros no vocábulo “gincana”; e 36% (N=31) de erros ortográficos na palavra “jiló”. Já nas palavras de maior frequência (barras verdes), ocorreram 2% (N=2) de erros na palavra “general”; 42% (N=33) de erro no vocábulo “jipe”; por fim, 29% (N = 21) de erros no vocábulo “sujeito”. Não houve erros na palavra “gigante”.

Aplicamos um teste Qui- quadrado para avaliar a diferença de erros ortográficos nas palavras de maior e menor frequência de ocorrência sem considerar o ano escolar. O teste confirmou que há diferença nos índices de erros ortográficos entre as palavras de menor e maior frequência de ocorrência (X^2-squared = 120.57, df = 7, p-value < 2.2e-16). Portanto, podemos afirmar que as maiores porcentagens de erros ortográficos, sem considerar o ano escolar, ocorrem em palavras de baixa frequência. Isso evidencia que a atuação da frequência de ocorrência pode depender de outros fatores, como o ano escolar, a relação fonema grafema, dentre outros não avaliados nesta tese.

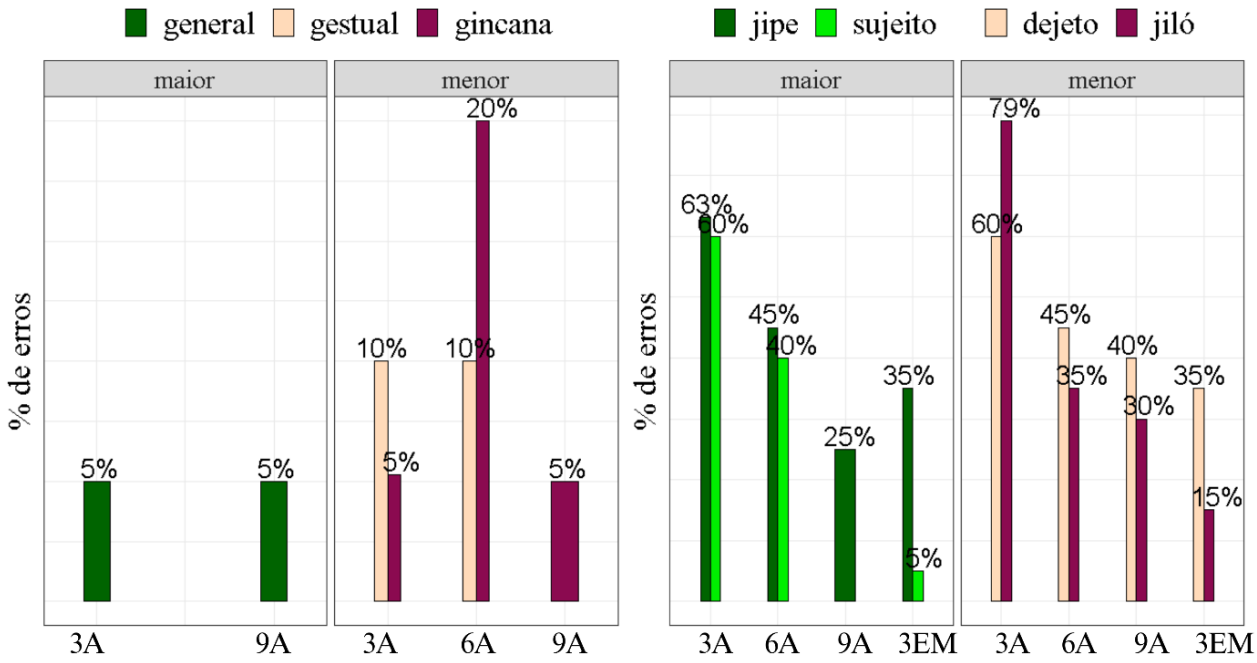
Além disso, ao observar o Gráfico 12, percebemos que os maiores índices de erros ortográficos ocorrem em palavras que são escritas com a letra <j>. Há a tendência de que os erros ortográficos dessas palavras sejam a escrita de <g> no lugar de <j>, como *degeto (“dejeto”), *giló (“jiló”); *gipe (“jipe”), *sugeito (“sujeito”). Mesmo em palavras com maior

frequência de ocorrência, como “jipe” e “sujeito”, os participantes tenderam a escrevê-las com <g> (maior frequência de tipo).

Ora, por que que mesmo em palavras de maior frequência, há a alta ocorrência de erros ortográficos? Um reposta para esta pergunta trata-se da alta diferença de frequência de tipo entre o <g> e <j>. O grafema <g> ocorre três vezes a mais que o <j>. Quando a frequência tipo é marcada, ou seja, há mais casos de palavras escritas com <g> do que com <j>, a atuação da frequência de tipo se torna maior do que a frequência de ocorrência. Portanto, mesmo em palavras com maior frequência de ocorrência, mas escritas com grafemas com menor frequência de tipo, a tendência é manter a maior frequência de tipo. Por isso, os dados do Gráfico 12 nos mostra alto índice de erros ortográficos em palavras escritas ortograficamente com <j>. Assim como nas pseudopalavras, discutidas no início do Capítulo, há uma preferência pelo <g> nas palavras reais.

Atrelado à frequência de ocorrência, parece que há a atuação da frequência de tipo no processo de escrita. Observaremos esta relação entre frequência de tipo e de ocorrência no gráfico a seguir.

Gráfico 13 - Frequência de tipo de <g> e <j> e frequência de ocorrência por ano escolar e por palavras com erros ortográficos



Fonte: Elaboração própria.

No gráfico 13, observaremos a relação entre frequência de tipo e de ocorrência por ano escolar. Podemos ler o Gráfico 10 da seguinte forma: no eixo X, há os anos escolares 3A (3º

ano do EF), 6A (6º ano do EF), 9A (9º ano do EF), 3EM (3ª série do EF) e, no eixo Y, a porcentagem de erros ortográficos em palavras escritas com <g> e <j> diante de <e> e <i>. Nos dois retângulos à esquerda, há o grupo de palavras escritas ortograficamente com <j> (menor frequência de tipo). Nos retângulos à direita, há o grupo de palavras escritas ortograficamente com <g> (maior frequência de tipo). Cada grupo de palavras está organizado em relação à frequência de ocorrência, a saber: as palavras nas cores verdes (verde escuro e claro) são de maior frequência de ocorrência; nas cores roxo e rosa claro, há as palavras de baixa frequência de ocorrência.

Ao observar as palavras escritas com <g> (maior frequência de tipo), os erros ortográficos ocorreram apenas na palavra "general" (maior frequência de ocorrência) no 3º ano e 9º ano do EF. Já nas palavras de menor frequência de ocorrência e escritas com <g>, os índices de erros aumentaram nos anos escolares 3º, 6º e 9º anos do EF. Além disso, aos observar as palavras escritas com <j> (menor frequência de tipo), há uma maior ocorrência de erros ortográficos, mesmo quando se trata de palavras de maior frequência de ocorrência. Portanto, podemos inferir que a frequência de ocorrência e a frequência de tipo podem ter uma diferente interação a depender da relação fonema-grafema. Por exemplo, nos dados do fonema /s/, como discutido na seção anterior, há menores índices de erros ortográficos em palavras de maior frequência de ocorrência, independente se são escritas ortograficamente com <c> (menor frequência tipo) ou com <s> (maior frequência de tipo).

7.1.2.3 Fonema /z/ e os grafemas <z> e <s>

Os grafemas <s> e <z>, entre vogais, concorrem para representar o fonema /z/. Há, por exemplo, as palavras "surpresa" e "beleza", "divisa" e "ojeriza". Considerando a comparação entre ambos, classificamos, nesta tese, o grafema <z> de menor frequência e o <s> de maior frequência, ambos no contexto descrito. Para a Tarefa 2 – palavras reais, selecionamos quatro palavras iniciadas com <z> e quatro iniciada com <s> no LexPorBR e LexPorBR – Infantil. Duas dessas quatros palavras possuem maior frequência de ocorrência e as outras duas, menor frequência de ocorrência. O Quadro 19 apresenta essas palavras, que estão organizadas em maior e menor frequência de ocorrência e em relação aos grafemas <s> e <z> no contexto descrito.

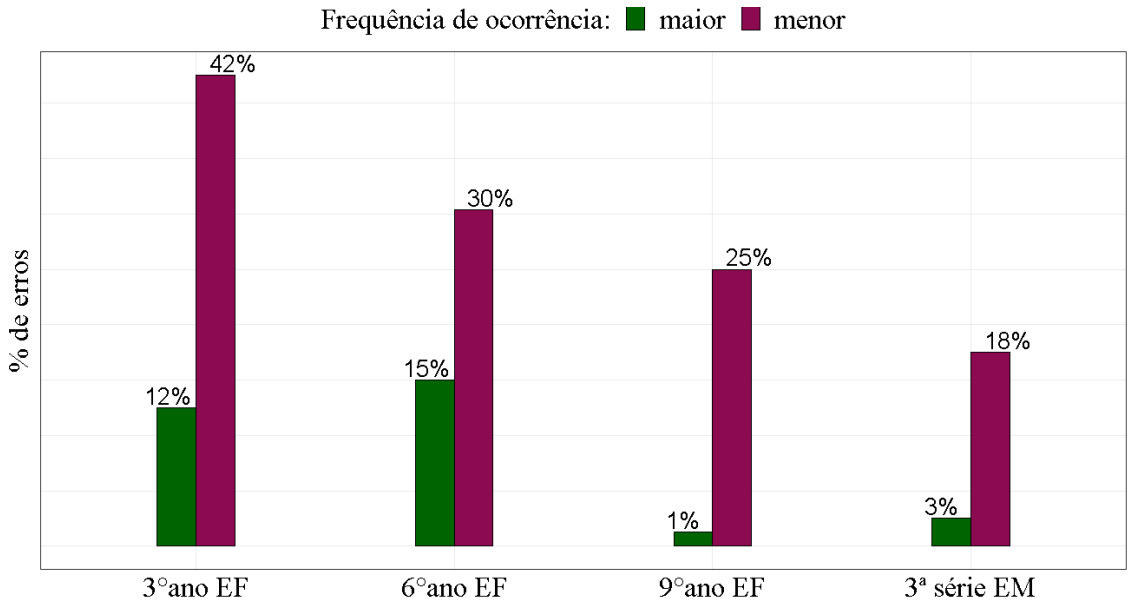
Quadro 19 – Palavras reais com os grafemas <s> e <z> que representam o fonema /z/

Frequência de ocorrência	Palavras			
	Grafema <s>		Grafema <z>	
Maior	surpresa	camisa	beleza	amizade
Menor	turquesa	divisa	leveza	ojeriza

Fonte: Elaboração própria.

Ao ouvirem os estímulos das palavras reais e observarem as imagens que remetiam a elas, os participantes as escreviam de acordo com o que eles julgassem ser a forma correta. Em relação a estes dados, filtramos apenas os erros ortográficos e os analisaremos. No Gráfico14, a seguir, há a porcentagem de erros ortográficos por ano escolar e pela frequência de ocorrência.

Gráfico 14 – Fonema /z/ e os erros na escrita das palavras reais com <s> e <z>



Fonte: Elaboração própria.

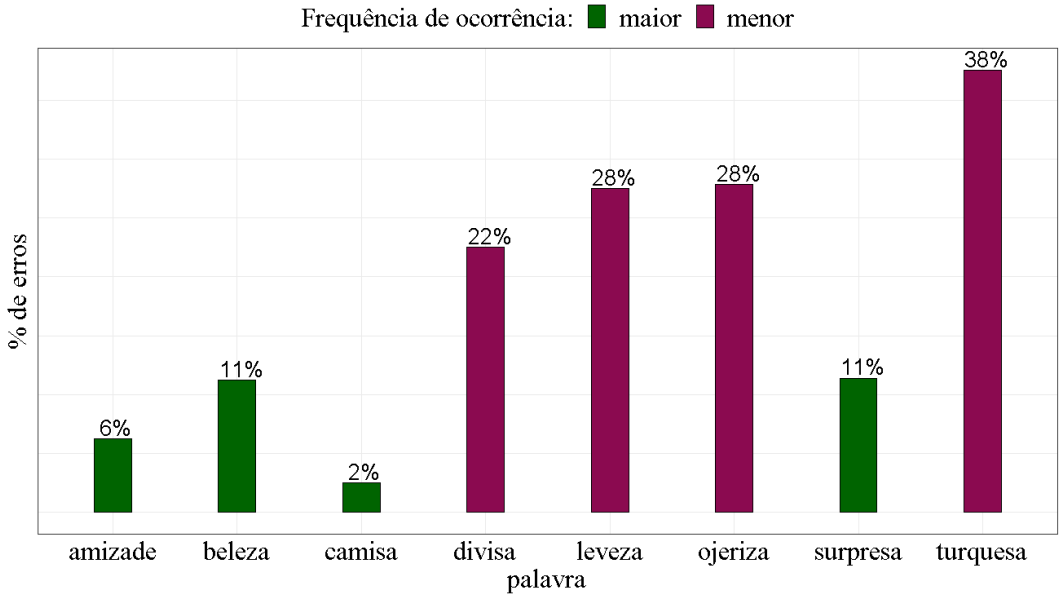
No eixo X, há os anos escolares, e o no eixo Y a porcentagem de erros ortográficos em palavras escritas com <s> ou <z> para representarem o fonema /z/. As barras de cor roxa são referentes aos erros ortográficos em palavras de menor frequência de ocorrência, já as barras de cor verde são erros em palavras de maior frequência de ocorrência. Nas palavras de maior frequência (cor verde escuro), os erros ortográficos ocorreram em 12% (N = 10) dos dados no 3º ano do EF; em 15% (N = 12) no 6º ano do EF; em 1% (N = 1) dos dados no 9º ano do EF; e em 3% (N = 2) dos dados na 3ª série do EM. Já nas palavras de menor de frequência de ocorrência (cor roxa), foram encontrados erros ortográficos em 42% (N = 34) dos dados no 3º

ano do EF; em30% (N = 2 4) dos dados no 6º ano; em 25% (N = 20) dos dados no 9º ano; e 18% (N = 14) dos dados na 3ª série do Ensino Médio.

Ao analisar o Gráfico 14, podemos observar que a taxa de erros ortográficos reduz em cada ano escolar. Além disso, é perceptível que as altas porcentagens de erros ortográficos ocorrem em palavras de menor frequência de ocorrência. Para confirmar se há diferença entre o índice de erros ortográficos em palavras com maior (cor verde escuro) e com menor frequência de ocorrência (cor roxa) em cada ano escolar, aplicamos um teste Qui-quadrado que mostrou que há diferença significativa entre estes dois grupos de palavras (X2- X-squared = 59.205, df = 7, p-value = 2.175e-10). Portanto, assim como nos dados relacionados ao fonema /s/ analisados anteriormente, os alunos são sensíveis à frequência de palavras e resolvem os problemas de escrita em palavras de maior frequência de ocorrência. Isso corrobora pesquisas de Santos e Befi-Lopes (2013) e de Ribeiro de Martins (2020), as quais pontuam que os alunos, em idade escolar, erram mais na escrita de palavras infrequentes. Portanto, ao olhar mais de perto os erros ortográficos, observamos que erros ocorrem em palavras de ortografia irregular, e dentro deste grupo, nas de menor frequência.

Além disso, é importante investigar em quais palavras ocorrem os erros ortográficos. Para isso, analisaremos o Gráfico 15, em que há a porcentagem de erros ortográficos no eixo Y, e por palavras no eixo X.

Gráfico 15 – Erros ortográficos por palavras com <s> e <z>

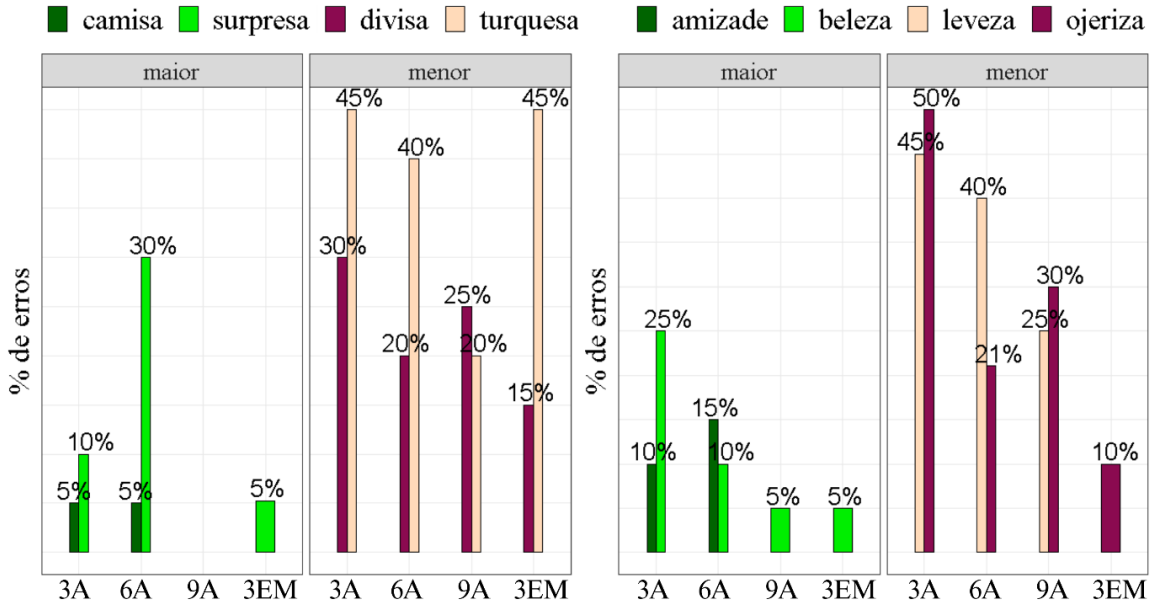


Fonte: Elaboração própria.

As barras de cor roxa são referentes aos erros ortográficos em palavras de menor frequência de ocorrência, já as barras de cor verde são erros em palavras de maior frequência de ocorrência. Nas palavras de menor frequência de ocorrência (barras roxas), ocorreram 22% (N= 18) de erros ortográficos no vocábulo “divisa”; 28% (N= 22) de erros na palavra “leveza”; 28% (N=22) de erros no vocábulo “ojeriza”; e 38% (N=30) de erros ortográficos na palavra “turquesa”. Já nas palavras de maior frequência de ocorrência (barras verdes), ocorreram 6% (N=5) de erros na palavra “amizade”; 11% (N=9) de erro no vocábulo “beleza”; 2% (N = 2) de erros no vocábulo “camisa”; e, por fim, 29% (N = 9) de erros no vocábulo “surpresa”.

Aplicamos um teste Qui-quadrado para avaliar a diferença de erros ortográficos nas palavras de maior e de menor frequência de ocorrência sem considerar o ano escolar. O teste confirmou que há diferença nos índices de erros ortográficos entre as palavras de menor e maior frequência de ocorrência (X^2-squared = 45.94, df = 7, p-value = 8.979e-8). Portanto, podemos afirmar que as maiores porcentagens de erros ortográficos, sem considerar o ano escolar, ocorrem em palavras de baixa frequência. Há a tendência de que os erros ortográficos dessas palavras sejam a escrita de <z> no lugar de <s>; ou de <s> no lugar de <z>, como *turqueza (“turquesa”), *levesa (“leveza”); *ojerisa (“ojeriza”), *diviza (divisa). Ressaltamos ainda que a escrita destas palavras pode ter influência da morfologia (sufixo -eza e -esa), a qual não foi testada nesta tese. Observaremos a relação entre frequência de tipo e de ocorrência no gráfico a seguir.

Gráfico 16 - Frequência de tipo de <s> e <z> e frequência de ocorrência por ano escolar e por palavras com erros ortográficos



Fonte: Elaboração própria.

Podemos ler o Gráfico 16 da seguinte forma: no eixo X, há os anos escolares 3A (3º ano do EF), 6A (6º ano do EF), 9A (9º ano do EF), 3EM (3ª série do EF) e o eixo Y a porcentagem de erros ortográficos em palavras escritas com <s> e <z> entre vogais. Nos dois retângulos à esquerda, há o grupo de palavras escritas ortograficamente com <z> (menor frequência de tipo). Nos retângulos à direita, há o grupo de palavras escritas ortograficamente com <s> (maior frequência de tipo). Cada grupo de palavras está organizado em relação à frequência de ocorrência: as palavras nas cores verdes (verde escuro e claro) são de maior frequência de ocorrência; nas cores roxo e rosa claro, há as palavras de baixa frequência de ocorrência. Ao observar o Gráfico 16, é perceptível que os maiores índices de erros ortográficos ocorreram em palavras de baixa frequência de ocorrência.

Na análise das pseudopalavras, observamos que os participantes da 3ª série do EM tenderam a escrever a letra <z> (de baixa frequência) em palavras com o som [z]. Nas análises das palavras reais, eles continuaram com esta tendência de escrever palavras com <z>, mesmo em casos de a ortografia ser com <s>, como em "turquesa" e "divisa". Isso pode ter acontecido porque os alunos deste ano tenderam a realizar uma relação direta entre o som [z] e o grafema <z>. Na análise da explicação oral dos sujeitos pesquisados (na seção da análise qualitativa), observaremos que as respostas dos sujeitos em relação à escolha da letra sempre passaram pela relação direta entre [z] e <z>.

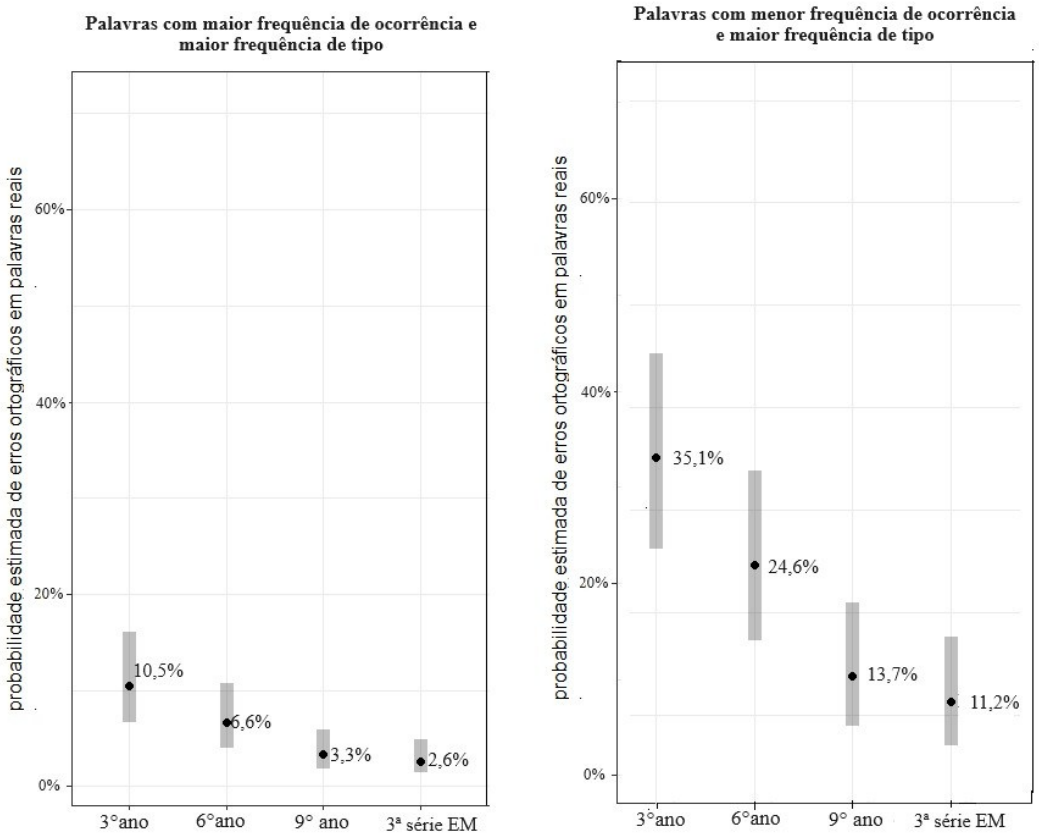
7.1.3 Discussão geral dos resultados frequência de ocorrência – palavras reais

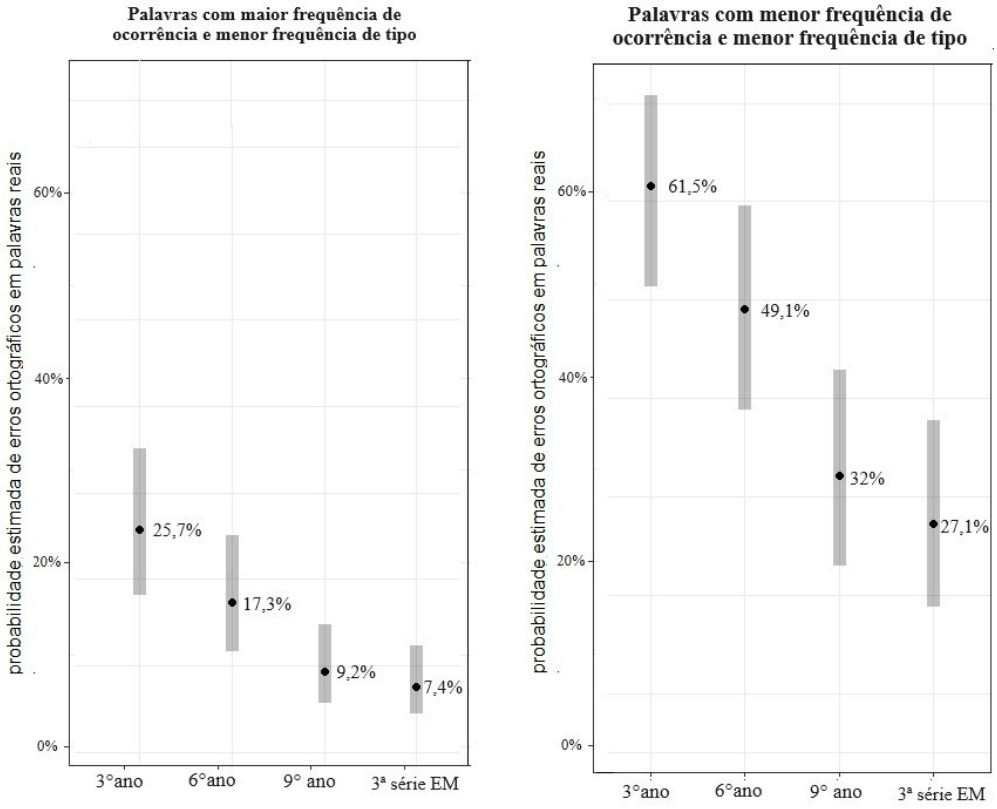
Nesta seção, investigamos se os três fatores pesquisados nesta tese — (1) ano escolar, a (2) relação fonema-grafema, a (3) frequência de tipo e a (4) frequência de ocorrência — influenciam na escrita de grafemas concorrentes irregulares. Para isso, um modelo linear generalizado misto[24] foi ajustado à amostra de dados das palavras reais. A função do modelo foi de estimar a probabilidade de ocorrência de erros ortográficos em interação com o ano escolar, a frequência de ocorrência e a frequência de tipo.

A variável dependente do modelo foi categórica: escrita das palavras (certo/errado). As variáveis independentes de efeito fixo do modelo foram: o ano escolar (3A, 6A, 9A e 3EM), a frequência de ocorrência (maior/menor) e a frequência de tipo (maior/menor). Além disso, o participante foi incluído ao modelo como variável independente de efeito aleatório. Uma comparação de modelos aninhados mostrou que o melhor modelo ajustado considera o índice de erro ortográfico em interação com o ano escolar, a frequência de ocorrência e a frequência de tipo (X^2 = 1685.3 77.307, p < 2.2e-16). Em outras palavras, o melhor modelo ajustado aos

dados indica que o índice de erro ortográfico é influenciado pelo ano escolar em interação à frequência de ocorrência e frequência de tipo. No Gráfico 17, há a probabilidade estimada pelo modelo em que o índice de erros ortográficos é explicado pelo ano escolar, pela frequência de ocorrência e pela frequência de tipo.

Gráfico 17 – Probabilidade estimada de erro ortográfico em palavras reais





Fonte: Elaboração própria.

O Gráfico 17 está organizado da seguinte forma: a probabilidade de erro ortográfico está no eixo Y. No eixo X, há os anos escolares. No primeiro retângulo à esquerda, observamos a probabilidade estimada de erros ortográficos em palavras com maior frequência de ocorrência e menor frequência de tipo. No primeiro retângulo à direita, há a probabilidade estimada de erros ortográficos em palavras com menor frequência de ocorrência e maior frequência de tipo. No terceiro retângulo, à esquerda, há a probabilidade estimada de erros ortográficos em palavras com maior frequência de ocorrência e menor frequência de tipo. Por fim, no quarto retângulo à direita, há a probabilidade estimada de erros ortográficos em palavras com menor frequência de ocorrência e menor frequência de tipo.

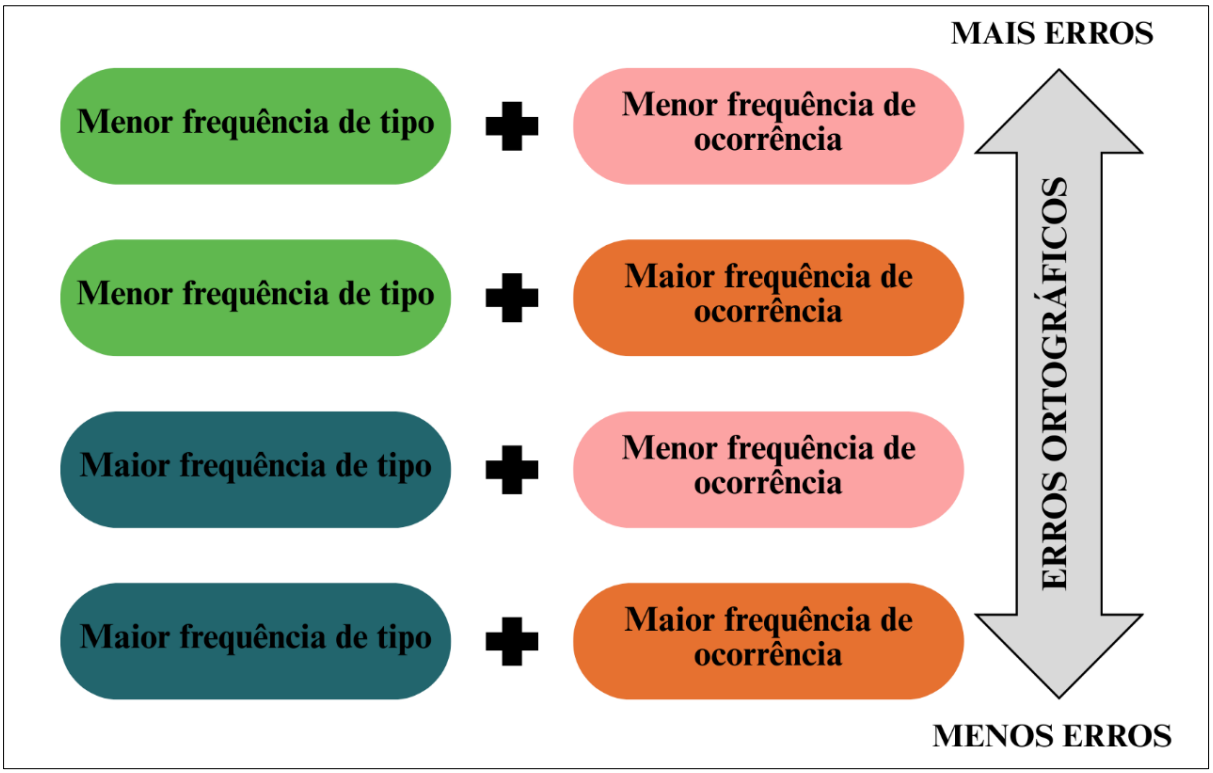
Podemos observar que quando há a interação entre maior frequência de tipo e maior frequência de ocorrência, a probabilidade de o erro ortográfico acontecer é pequena. Por exemplo, no 3º ano, a probabilidade é de 10,5% e de 2,6% na 3ª série do Ensino Médio. Ao passo que a probabilidade estimada de acontecer o erro ortográfico é alta quando há a interação de menor frequência de tipo e menor frequência de ocorrência. Por exemplo, a probabilidade de errar é 61,5% no 3º ano do EF e de 27,1% na 3ª série do Ensino Médio. Além disso, nas palavras de menor frequência de ocorrência e maior frequência de tipo, ou de maior frequência

de ocorrência e menor frequência de tipo, os dois retângulos do meio, a possibilidade de erro pode variar ao comprar com o último e o primeiro retângulo.

Vale ressaltar ainda que, além da probabilidade estimada pelo modelo, o Gráfico 17 nos mostra o intervalo de confiança. Segundo Oushiro (2017), o intervalo de confiança "estabelece um valor mínimo e um valor máximo em que se calcula estar o verdadeiro parâmetro da população". Em outras palavras, o intervalo de confiança calcula as incertezas relacionadas à probabilidade estimada do modelo. Quanto menor o intervalo de confiança, maior a previsibilidade do modelo. Ao observarmos a sombra cinza no Gráfico 17, podemos perceber que, mesmo considerando os intervalos de confiança, as chances de ocorrer o erro ortográfico na interação maior frequência de ocorrência e maior frequência de tipo é baixa. Por exemplo, a probabilidade de um aluno errar a escrita de "gigante" é baixa, pois esta palavra é de maior frequência de ocorrência e escrita com <g> de maior frequência de tipo.

Nesta seção, analisamos probabilidade estimada do modelo em relação ao erro ortográfico em cada ano escolar. Com os resultados do modelo, podemos concluir que há uma interação entre frequência de tipo e de ocorrência. O aluno, em seu percurso de aprendizagem, pode ser sensível tanto à frequência de um padrão específico, quanto à frequência de uma palavra. Este resultado é condizente com a perspectiva da aprendizagem estatística proposta pela IMP (Treiman; Kessler, 2014). No entanto, é importante considerar que a frequência, sozinha, não influencia o processo de escrita de grafemas concorrentes irregulares. Ela precisa ser considerada em relação ao ano escolar e ao tipo de frequência, se é de tipo ou de ocorrência. Além disso, este aprendizado pode interagir com outros padrões emergentes na língua, bem como outros fatores, como os padrões semânticos e morfológicos, os quais não foram testados nesta tese. A partir destes resultados, podemos realizar as seguintes generalizações.

Figura 11 – Generalizações das relações entre frequência de tipo e frequência de ocorrência



Fonte: Elaboração própria.

À esquerda, nas cores verde e azul, há a frequência de tipo; à direita, nas cores rosa e laranja, a frequência de ocorrência. A seta cinza indica a gradação dos erros ortográficos em função da frequência de tipo e de ocorrência. Segundo Bybee (2016), essa gradação, observada em categorias linguísticas, é conceituada como gradiência. Entre os extremos da seta cinza (maior e menor quantidade de erros ortográficos), observa-se uma gradiência que pode estar relacionada a múltiplos fatores que influenciam a escrita ortográfica. De modo geral, quanto menor a frequência de tipo e de ocorrência, maior tende a ser o número de erros ortográficos; por outro lado, quanto maiores essas frequências, menor pode ser a incidência de erros. Ao identificar os padrões e as palavras com maior ocorrência de erros ortográficos em sala de aula, o professor pode propor atividades reflexivas e sistematizadas, voltadas para as principais dificuldades de escrita dos alunos.

7.2 Análise qualitativa dos dados: da análise à discussão dos resultados

Relembrando, o objetivo geral desta tese é investigar a influência do ano escolar, do padrão ortográfico e a frequência de tipo e de ocorrência na escrita de grafemas concorrentes irregulares de alunos de uma escola de Belo Horizonte – MG. Como um dos caminhos para

atingir ao objetivo, optamos por, além da análise estatística realizada anteriormente, investigar, qualitativamente, o que os participantes desta pesquisa têm a dizer sobre o que eles escreveram na "Tarefa 1 – Pseudopalavras". Durante este experimento, os estudantes foram questionados sobre o motivo que os levaram a escolher determinado grafema para escrever algumas das pseudopalavras que eles ouviram. No momento da escrita das pseudopalavras, a pesquisadora perguntava: *"Por que você escreveu com determinada letra? O que você pensou sobre?"*. Estas perguntas foram realizadas ao final da escrita pelo participante de cada uma das seguintes psedopalavras: ['silʊ], ['sevʊ], ['ʒepʊ], ['ʒikʊ], ['bezə], ['nizə], ['tezə], ['dʒizə]. O objetivo foi captar diretamente o que o aluno pensou, sem influenciar suas respostas posteriores. Esta parte do experimento foi gravada e os áudios transcritos para um arquivo do Word. O anonimato dos participantes foi mantido e as identificações utilizadas nesta seção são apenas códigos para facilitar a discussão dos resultados.

Além disso, é importante ressaltar, como explicado na metodologia desta tese, que a pesquisadora respeitou os estudantes e não insistia para obter respostas. Por exemplo, os alunos do 3º ano do Ensino Fundamental I estavam tímidos e tinham medo de errar, mesmo com a orientação de que não seriam avaliados. Por outro lado, houve também a situação com o 9º ano do Ensino Fundamental II, cujos alunos não estavam interessados na pesquisa. As perguntas foram realizadas com alunos que se demonstraram abertos a elas, que não estavam tímidos ou desinteressados.

A seguinte pergunta guiou o desenvolvimento desta parte da pesquisa: "o que os alunos pensam ao ter que optar por grafemas que concorrem para representar um som?" Neste contexto, o objetivo desta seção é analisar, qualitativamente, os dados coletados a partir destes questionamentos e demonstrar os múltiplos padrões envolvidos no processo da escrita ortográfica de grafemas concorrentes irregulares. Nas subseções seguintes, os dados relacionados a cada um dos fonemas e seus grafemas são apresentados e discutidos.

7.2.1 Fonema /s/ e os grafemas <c> e <s>

O fonema /s/, em contexto de início de palavras diante de <e> e <i>, pode ser representado, na escrita, pelos grafemas <c> e <s>. Estes dois grafemas concorrem para representar o /s/ neste contexto, ou seja, não há nenhuma regra que regule a escrita de <c> ou <s> em início de palavras diante de <e> e <i>. Há casos, por exemplo, de "cebola" e "segredo", "cidade" e "sigilo". O questionamento foi realizado em duas das pseudopalavras iniciadas com o som [s]: ['sevʊ], ['silʊ]. Além das respostas relacionadas a ['sevʊ] e ['silʊ], trouxemos dois

exemplos de respostas relacionadas ao estímulo ['setʊ]. Esperávamos que os participantes escrevessem as pseudopalavras com o grafema <c> ou o <s>. Quando a escrita era realizada, a pesquisadora indagava aos estudantes: *"Por que você escreveu com <letra escolhida pelo participante>? O que você pensou sobre?"*.

No Quadro 20, a seguir, apresentamos as transcrições das respostas de alguns sujeitos. Na coluna "Estímulo", há as pseudopalavras que eles escutaram, seguida pelo código do participante e o ano escolar. Na coluna nomeada "Escrita", há o que o aluno escreveu. Já na última coluna, há a transcrição da resposta à pergunta realizada pela pesquisadora. As transcrições foram realizadas em consonância ao que foi dito, por isso, pode haver palavras ou construções de orações que não seguem regras gramaticais.

Quadro 20 - Transcrição das respostas dos participantes sobre a motivação da escolha de determinada letra para a escrita de pseudopalavras iniciadas com o som [s]

Estímulo	Participante	Ano escolar	Escrita	Resposta
['sevʊ]	WS6A	6º ano	sevo	Por causa que bem no comecinho tinha o som de 's'
	MP6A	6º ano	sevo	Escutei [sevʊ] (alongamento na produção do [s])
	MM6A	6º ano	sevo	Eu ouvi um som de 's'
	FP6A	6º ano	cevo	Fala 'c'… letra 'c'
	CS6A	6º ano	cevo	De cevada
	BR6A	6º ano	sevo	O 's' eu achei que está mais forte
	AN6A	6º ano	sevo	Veio um som de 'z', só que não é 'z' no sentido
	DM3S	3ª série EM	cevo	Pela tonalidade
	EP3S	3ª série EM	sevo	O 's' e o 'e' no início é mais puxado do que com o 'c'.
	GC3S	3ª série EM	sevô	Oh… esse eu olhei mais o desenho e a voz. Pelo desenho, parece mais pelo Pokémon. E o nome dos pokémons não têm muitos que comecem com 'c', a maioria é com 's'.
	MD3S	3ª série EM	sevo	Por causa que começa com 's' [...]. Se fosse com z, seria 'ze', e é 'se'
	RA3S	3ª série EM	sevo	Eu acho que estou acostumada a escrever com 's'
	TO3S	3ª série EM	cevo	Acho que com 's' ficaria estranho
	VB3S	3ª série EM	sevo	Porque faz som de "servo"
	YS3S	3ª série EM	sevo	Eu estava pensando [...], aí eu optei pelo 's'. Eu pensei no 'c' também, mas ia dar som de 'ca'
['silʊ]	WS6A	6º ano	silo	Tipo assim, o som dele (do 's') é mais fino, e do 'z' é mais grave

Estímulo	Participante	Ano escolar	Escrita	Resposta
	NB6A	6º ano	silo	Eu achei que era do ‘s’
	MP6A	6º ano	selo	Não tem ‘ç’ no início da palavra, quando é ‘c’ […] aqui ia ficar…sei lá… “selo”? Dá pra entender um pouco
	CS6A	6º ano	cilo	Esse aí é homem papelão pra mim […]. Cilo […] pensei em ‘c’ de “reciclar”, e ele é de papelão […], ele é reciclado
	AN6A	6º ano	silo	No significado está com som de ‘s’, não de ‘z’
	EP3S	3ª série EM	selo	Essa palavra ficou parecendo com uma que existe […], “selo”
	GC3S	3ª série EM	cilo	“Cilo” [...] por causa do som, do ‘ci’. O ‘s’ som não tem o mesmo som do c separado
	RA3S	3ª série EM	cilo	Acho que se fosse com ‘s’ ia ficar com som de ‘z’
['setʊ]	CS6A	6º ano	ceto	Com ‘c’ de […] tipo [...] “acetona”
	CF6A	6º ano	ceto	Parece que tem som de ‘c’

Fonte: *Corpus* da pesquisa.

A oportunidade de perguntar ao aluno o motivo que o levou a escrever determinada a letra revela como o raciocínio durante o ato de escrever pode seguir caminhos nunca imaginados pelo pesquisador. Além disso, essa oportunidade também visava a dar voz ao protagonista da própria escrita, os sujeitos da pesquisa. Portanto, observar o percurso de aprendizagem dos alunos é uma oportunidade para um linguista que estuda o processo de aprendizagem da escrita ortográfica.

Ao analisar as respostas dos alunos no Quadro 20, podemos notar que o caminho da escrita é múltiplo e pode ser motivado por diversos conhecimentos. Há casos, por exemplo, em que o aluno partiu da relação entre letra-som para escrever a pseudopalavra escutada. Ao ouvir o estímulo ['sevʊ], participantes disseram que escolheram o <s> porque o som era de [s], como WS6A, MP6A e MM6A do 6º ano do EF. Além disso, também pelo caminho do som, estudantes comentaram que escolheram o grafema <s> por ele não ter o som de [k], como relatado pelo aluno YS3S da 3ª série EM. Essa resposta indica que o sujeito apenas observou a relação direta entre o som [k] e o grafema <c>, mas não considerou o contexto e demais relações ortográficas do fonema /s/ e o grafema <c> diante de <e> e <i> em início de palavra. O participante FP6A, do 6º ano, escreveu “cevo” e justificou a escolha de <c> por ter ouvido “letra ‘c’”. Este aluno pensou no nome da letra <c>, lê-se “cê”.

Por outro lado, o GC3S, da 3ª série do EM, ao ouvir ['sevʊ], relacionou a escrita de <s> à imagem observada do *Digimon*. É importante relembrar que os participantes, durante a Tarefa 1 – pseudopalavras, viram desenhos de *Digimons*, criaturas de uma franquia homônima de mídia japonesa e ouviram os apelidos que a pesquisadora criou para eles. Então, o GC3S justificou a sua escolha por <s> por não ser comuns nomes de pokémons que começam com <c>, criaturas que se assemelham aos *Digimons* mas pertencem a outra franquia de mídia japonesa, também homônima. Além disso, a resposta do participante TO3S, da 3ª série do EM, chama a atenção ao afirmar que está acostumado a escrever com <s>. Esta resposta nos leva a refletir sobre a sensibilidade do estudante à frequência do grafema. De fato, como discutido nas seções anteriores, o grafema <s> em início de palavra diante de <e> e <i> ocorre mais do que o <c>. Essa sensibilidade à frequência pode ser um dos caminhos que explique esta resposta do estudante. Ainda no estímulo ['sevʊ], participantes relacionaram a escrita das pseudopalavras a palavras reais. Por exemplo, o CS6A justificou sua decisão por "cevo" pela palavra "cevada". Já o VB3S escreveu "sevo" por este ter o mesmo som que "servo".

Em relação ao estímulo ['silʊ], os participantes, ao serem questionados sobre o motivo da escrita de determinadas letras, utilizaram caminhos diversos, como o apoio à ortografia de palavras reais e à relação som e grafema. Por exemplo, o CS6A, do 6º ano, lembrou da palavra "reciclar" ao ver a imagem do "Digimon" do estímulo ['silʊ], o qual o remeteu a uma caixa de papelão. O aluno afirmou que o desenho, por lembrar um "homem papelão", o fez escrever "cilo", com <c>, por relacionar ao <c> de reciclar. Essa resposta mostra-nos que o processo de escrita pode seguir caminhos diversos.

Além do CS6A, o EP3S, da 3ª série do EM, também relacionou a escrita a uma palavra real. Ao ouvir ['silʊ], o participante lembrou de "selo", por isso, optou por escrever a letra <s>. Quanto a isto, é interessante observar que o aluno buscou uma palavra com esquema segmental semelhante e ainda seguiu uma tendência subjacente à ocorrência de [i] e [u] que podem alternar com [e] e [i]. Embora seja importante deixar claro que esta alternância não é verificada em contexto tônico, como o da palavra da "cilo", a possibilidade de <e> e <i> alternarem, mesmo que em outros contextos, abre as portas para esta possibilidade.

O RA3S, da 3ª série do EM, optou por escrever "cilo" com <c>, e não <s>, pois, com este último ficaria com o som [z]. Interessante observar que o aluno supôs uma relação direta entre fonema e grafema, de que o <s> representa diretamente o /z/. O aluno não considera as possibilidades de o fonema /s/ também ser representado por <s> em alguns contextos. Por outro lado, os participantes WS6A e AN6A, ambos do 6º ano, escreveram a letra <s> com a justificativa de que ouviram o som [s], e não o [z]. Estes alunos consideraram apenas a relação

fonema e grafema, sem pensar no contexto. Essa relação direta fonema – grafema foi notável na escrita de alunos nos dados com o fonema /z/ e, principalmente, na 3ª série do EM.

Os participantes CS6A e CF6A, sem serem questionados, explicaram o motivo de usarem a o grafema <c> na escrita de tal pseudopalavra. O CS6A relacionou a escrita à palavra "acetona", já o CF6A guiou-se pela relação com o som e colocou a letra <c> com a justificativa de ter ouvido a letra "cê".

Ao longo dessa seção, analisamos as respostas dos alunos ao questionamento do motivo que os levaram a escrever determinadas letras para os estímulos ['sevʊ], ['silʊ] e ['setʊ]. A partir das análises, foi possível identificar que os participantes seguem caminhos diversos para a escrita de uma palavra, como a relação com outros vocábulos, a relação direta entre som e letra, as relações contextuais entre os fonemas /s/ e /z/ e os grafemas <s> e <z>, os fonemas /k/ e /s/ e o grafema <c>, até mesmo a relação gráfico semântica entre a pseudopalavra e a imagem do Digimon observada. Estes caminhos diversos são condizentes à concepção da integração de múltiplos padrões no processo da aprendizagem da escrita ortográfica proposto pela IMP (Treiman, Kessler, 2014).

7.2.2 Fonema /ʒ/ e os grafemas <g> e <j>

O fonema /ʒ/, diante de <e> e <i>, pode ser representado, na escrita, pelos grafemas <g> e <j>. Estes dois grafemas concorrem para representar o / ʒ / neste contexto, ou seja, não há nenhuma regra que regule a escrita de <g> ou <j> diante de <e> e <i>. Há casos, por exemplo, de "geração" e "jejum", "gigante" e "jipe".

Para a coleta de dados do fonema /ʒ/ e os grafemas <g> e <j>, o questionamento aos alunos do 3º ano EFI, 6º ano do EFII e 3ª série do EM foi realizado em duas das pseudopalavras iniciadas com o som [ʒ], a saber: ['ʒikʊ], ['ʒepʊ]. Além das respostas relacionadas a ['ʒikʊ] e ['ʒepʊ], apresentamos a resposta em relação aos estímulos ['ʒevʊ] e ['ʒetʊ]. Quando a escrita dos estímulos era realizada, a pesquisadora indagava aos sujeitos com as perguntas: *"Por que você escreveu com <letra escolhida pelo participante>? O que você pensou sobre?"*.

No quadro a seguir, há as respostas transcritas. Na coluna "Estímulo", há as pseudopalavras que eles escutaram, seguida pelo código do participante e o ano escolar. Na coluna nomeada "Escrita", há o que o aluno escreveu. Já na última coluna, há a transcrição da resposta à pergunta realizada pela pesquisadora. As transcrições foram realizadas em consonância ao que foi dito pelos participantes, portanto, pode haver palavras ou construções de orações que não seguem regras gramaticais.

Quadro 21 - Transcrição das respostas dos participantes sobre a motivação da escolha de determinada letra para a escrita de pseudopalavras iniciadas com o som [ʒ]

Estímulo	Participante	Ano escolar	Escrita	Resposta
['ʒikʊ]	DA3A	3º ano EFI	jico	O 'j' tem o mesmo som do 'g'. Se eu colocar 'g' ou 'j' vai ficar do mesmo jeito
	WS6A	6º ano	jico	O 'g' é um som um pouco mais aberto e o 'j' mais fechado
	GC3S	3ª série EM	jico	"Jico" por causa do jeito que falou [...] "jico". Eu acho que é com 'j', porque com 'g' não ia fazer muito sentido
	CS6A	6º ano	jico	'j' de jesus que bicho fofo
['ʒepʊ]	CS6A	6º ano	jepo	eu botei com J por causa que eu achei esse bichinho muito fofo aí eu pensei: Jesus que bicho fofo
	WS6A	6º ano	gepo	Não sei explicar, mas o 'g' e o 'j', um tem o som mais aberto e o outro mais fechado
	NB6A	6º ano	gepo	Na primeira parte, ela fala 'g'
	MP6A	6º ano	jepo	Tipo assim, igual, tipo, tem nome "Jhonatan" e "Gustavo". Aí dá pra você saber o significado
	MM6A	6º ano	gepo	Eu ouvi o som de 'g'
	MA6A	6º ano	jepo	"Jebo" […] tipo assim [...]. Não é muito bom o 'g' e não é muito bom o 'j'
	CF6A	6º ano	jepo	Ah... parece que tem som de 'j'
	BR6A	6º ano	gepo	Eu acho assim… que com 'g' fica mais forte
	KC3S	3ª série EM	gepo	Por causa da audição, me remeteu ao 'g'
	MD3S	3ª série EM	gepo	Por causa que [...] pode ser com 'j'. Só que optei por 'g'
	RA3S	3ª série EM	gepo	Combina mais com 'g', com 'j' ia ficar estranho
	TO3S	3ª série EM	gepo	Faz mais sentido para mim ser com 'g'
	YS3S	3ª série EM	gepo	"Gepo'… soou com som de 'g'
['ʒidʃɪ]	MD3S	3ª série EM	jidi	Porque o 'g' expressa uma expressão e o 'j' outra. Tipo jade. [pesquisadora: e com 'g' seria como?] Gade
['ʒevʊ]	CS6A	6º ano	gevo	Com 'g' de "gesso"
['ʒetʊ]	CS6A	6º ano	jeto	O meu maior problema é saber se é com 'g' ou com 'j'. Neste caso, eu escutei 'j'.

Fonte: *Corpus* da pesquisa.

A partir das respostas dos participantes em relação ao questionamento da escrita de ['ʒikʊ] e ['ʒepʊ], transcritas na última coluna do Quadro 21 acima, podemos observar as diversas

relações que eles fazem no momento da escrita de não-palavras. Há respostas relacionadas a irregularidades ortográficas, a sons de grafemas, à ortografia de palavras reais, a nomes de letras, a relações ortográficas contextuais.

A resposta o DA3A em relação à escrita de <jico> para o estímulo ['ʒikʊ] apresenta um alto conhecimento ortográfico, principalmente nos casos de irregularidades ortográficas, ainda que no 3º ano do EFI, ano que inicia o ensino sistemático da ortografia. O aluno respondeu que o <j> tem o mesmo som de <g> e ao escolher um ou outro, o som continuaria o mesmo. De acordo com a BNCC (Brasil, 2017), o ensino sistemático da ortografia é iniciado no 3º ano a partir das relações ortográficas regulares e contextuais. O ensino das irregularidades apenas é desenvolvido no 5º ano. Mesmo não tendo recebido o ensino sistemático das irregularidades ortográficas, o DA3A já apresenta o conhecimento de que há casos em que várias letras podem representar o mesmo som. Isso pode ser indícios da aprendizagem estatística, pois o aluno conseguiu mapear padrões e chegar à generalização de que <g> e <j> representam o mesmo som.

Ainda no estímulo ['ʒikʊ], o CS6A escolheu o <j> ao relacionar ao <j> da palavra "jesus" ao de "jico". O aluno considerou "fofo" o *Digmon* apresentado e utilizou uma expressão mineira de usar o nome "Jesus" para se referir a algo expressivo. No caso, o estudante disse "jesus, que bicho fofo" e escreveu "jico" relacionando ao <j> de "Jesus". Este mesmo participante utiliza esta resposta para justificar a escrita de <jepo>. De acordo com o aluno, ele colocou com <j> por ter achado "esse bichinho muito fofo", por isso ele pensou: "jesus que bicho fofo".

Em relação ao estímulo ['ʒepʊ], o MP6A (6º ano) abordou um interessante aspecto da relação regular contextual entre fonema e grafema. Por exemplo, o participante escreveu <jepo> e utilizou a justificativa da escolha por <j> por haver, na língua, casos como "Jhonatan" e "Gustavo". Dessa forma, por ter ouvido o som [ʒ], optou por <j>, e não por <g> que pode representar o som [g], como o segundo nome por ele citado. Por outro lado, seguindo apenas a relação som e letra, os participantes NB6Ae MM6A do 6º ano e o KC3S e YS3S da 3ª série do EM afirmaram que escreveram <g> por terem ouvido [g]. Esses participantes podem ter relacionado a escrita de <g> ao nome dessa letra, lê-se "gê". Ao ouvirem ['ʒepʊ], eles podem ter, automaticamente, lembrado do nome da letra "gê".

Além disso, é importante observar das muitas respostas que argumentavam que a escrita de <g> ocorreu, pois com <j> "ficaria estranho", como para o RA3S da 3ª série do EM; com <g> "faz mais sentido", como para o TO3S; "com 'g' fica mais forte" para o BR6A. Essas respostas podem ser evidência do pensamento de que escrever um grafema que é mais

frequente, como o <g>, é o correto a se fazer. Mesmo não tendo utilizado o termo “frequente” em suas respostas, podemos inferir que eles se referiram à frequência do grafema, pois, como discutido na seção anterior, os alunos de todos os anos escolares têm preferência pela escrita de <g>, que ocorre três vezes mais que o <j> em início de palavras.

O participante CS6A se entusiasmava em cada um dos estímulos que ouvia e com os *Digmons* que apareciam. Ele justificou a escolha das letras em cada um dos itens, mesmo sem o questionamento da pesquisadora. Por isso, além de ['ʒikʊ] e ['ʒepʊ], também apresentamos a resposta do estudante em relação aos estímulos ['ʒevʊ] e ['ʒetʊ]. Na primeira pseudopalavra, o sujeito justificou a escrita de “gevo” com o “g” de “gesso”, ou seja, ele recorreu à ortografia de uma palavra real para escrever ['ʒevʊ]. Já na segunda pseudopalavra, o aluno destacou que um dos seus problemas na escrita é saber se usa o <g> ou o <j>. Isso é o problema de muitos, pois, com discutimos ao longo desta tese, não há regra para a escrita de <g> e <j>, precedendo a vogal anterior, ou seja, os dois grafemas concorrem para representar o fonema /ʒ/ neste contexto.

Assim como discutido na seção anterior das respostas relacionadas ao fonema /s/, nesta seção foi perceptível que os participantes seguem caminhos diversos para a escrita de uma palavra, como a relação a outros vocábulos, a relação direta entre som e letra, as relações contextuais entre os fonemas /ʒ/ e /g/ e os grafemas <g> e <j>, até mesmo a relação gráfica entre a pseudopalavra e a imagem do *Digimon* observada. Estes caminhos diversos são condizentes à concepção da integração de múltiplos padrões no processo da aprendizagem da escrita ortográfica proposto pela IMP (Treiman, Kessler, 2014), a qual postula que a aprendizagem da ortografia é motivada por múltiplos padrões, como o fonético, fonológico, gráfico, semântico, morfológico, entre outros.

7.2.3 Fonema /z/ e os grafemas <z> e <s>

O fonema /z/, entre vogais, pode ser representado, na escrita, pelos grafemas <s> e <z>. Estes dois grafemas concorrem para representar o /z/ neste contexto, ou seja, não há nenhuma regra que regule a escrita de <s> ou <z> entre vogais. Há casos, por exemplo, de “surpresa” e “leveza”, “divisa” e “ojeriza”. Para a coleta de dados do fonema /z/ e os grafemas <s> e <z>, o questionamento aos alunos foi realizado em quatro das pseudopalavras com o fonema /z/, escolhidas aleatoriamente: ['bezə], ['nizə], ['tezə], ['dʒizə].

No Quadro 22, a seguir, há as respostas transcritas. Na coluna “Estímulo”, há as pseudopalavras que eles escutaram, seguida pelo código do participante e o ano escolar. Na

coluna nomeada “Escrita”, há o que o aluno escreveu. Já na última coluna, há a transcrição da resposta do participante à pergunta realizada pela pesquisadora. As transcrições foram realizadas em consonância ao que foi dito pelos participantes, por isso, pode haver palavras ou construções de orações que não seguem regras gramaticais.

Quadro 22 - Transcrição das respostas dos participantes sobre a motivação da escolha de determinada letra para a escrita de pseudopalavras com o som [z]

Estímulo	Participante	Ano escolar	Escrita	Resposta
['bezə]	DA3A	3º ano EFI	pezar	O ‘s’ tem o mesmo som que o ‘z’
	MP6A	6º ano	beza	Porque tá com um som meio de ‘z’
	EB6A	6º ano	beza	Na minha lógica, o ‘s’ a gente só usa quando é palavra mais fraca ou quando é o ‘ss’. Aí quando é palavra assim, eu acho que é com ‘z’, porque é uma palavra mais forte
	DB6A	6º ano	besa	São poucas palavras que tem com ‘z’
	EP3S	3ª série EM	peza	Pela fala mesmo, porque o “peza”, o ‘za’ saiu mais forte do que como se fosse com ‘sa'. Quando fala ‘sa’ costuma ser falado de uma forma suave
	GC3S	3ª série EM	besa	O barulho de um e do outro são idênticos, para distinguir é muito complicado
	MB3S	3ª série EM	besa	Está próximo de “Teresa”
['nizə]	CS6A	6º ano	niza	Sei lá... porque eu lembrei de “Luiza”. “Luiza” costuma ser com ‘z’
	BR6A	6º ano	niza	O ‘s’ é muito forte, e só um (‘s’) não fica tão certo em um nome assim
	KC3S	3ª série EM	miesa	O som que eu ouvia remetia ao ‘z’, ‘zá’. Eu relacionei com o ‘s’
	MD3S	3ª série EM	neza	Por causa das expressões, por causa que ‘z’ e ‘s’ têm as mesmas expressões
	TO3S	3ª série EM	nyza	Acho que faz mais sentido, porque são personagens, né!?
	VB3S	3ª série EM	niza	É “niza”, puxa para o ‘z’
	YS3S	3ª série EM	niza	Acho que é com ‘z’, “niza”. Se fosse com ‘s’ ia ser “ssa” [referindo-se ao som [s]
['tezə]	WS6A	6º ano	teza	Por causa que tem o som mais um pouquinho grave
	MY6A	6º ano	tesa	Tem mais som de ‘s’
	MP6A	6º ano	teza	Dá para diferenciar o ‘s’ e o ‘z’… não sei … o ‘z’ puxa mais
	MA6A	6º ano	tesa	Eu lembro de “peso”, “peso” é com ‘s’
	FP6A	6º ano	tesa	As vezes muda também, tipo o ‘z’ e o ‘s’, mas às vezes são iguais
	EM6A	6º ano	tesa	“Tesa” uai… eu lembrei de “peso”, “peso” é com s

Estímulo	Participante	Ano escolar	Escrita	Resposta
	CS6A	6º ano	tesa	"Tesa", eu pensei, que errado. Se colocar 'o' e um til no 'a', ficaria muito errado (risos ao lembrar da palavra "tesão")
	EP3S	3ª série EM	teza	Quando é 'z', a fala é mais forte e mais alta, ao invés do 's'
	MB3S	3ª série EM	tesa	Me lembra 'Teresa'
	MD3S	3ª série EM	teza	Eu acho que também pode ser com 's', mas eu optei por 'z'. Por que ia parecer coisa inapropriada, e eu não quero… te respeito [se referendo à palavra tesão]
['dʒizə]	BG3A	3º ano EFI	diza	O 's' não estava dando o som de 'z'
	GC3S	3ª série EM	diza	"Diza", porque essa puxa mais para a parte do 'z', de "za", não puxa muito para a parte do 's'
	MB3S	3ª série EM	diza	Me lembra "dizer"
	MD3S	3ª série EM	guiza	Por causa da expressão 'Zê', com "s" seria "guissa" [se referindo ao som [s], com 'z' é "guiza"

Fonte: *Corpus* da pesquisa.

A partir das respostas em relação ao questionamento da escrita de ['bezə], ['nizə], ['tezə] e ['dʒizə] na última coluna do quadro, podemos observar as diversas relações que eles fazem no momento da escrita de não-palavras. Há respostas relacionadas à ortografia de palavras reais, a relações ortográficas regulares contextuais, a irregularidades na ortografia e a relação direta entre fonema e grafema.

Há participantes que relacionaram a escrita das paseudopalavras à ortografia das palavras reais. O MB3S, da 3ª série, recorreu à palavra "Teresa" para a escrita de "besa" e "tesa". O CS6A, do 6º ano, recorreu ao nome "Luiza" para a escrita de "niza". Dois participantes do 6º ano, MA6A e EM6A recorreram à palavra "peso" para a escrita de "tesa". No estímulo ['dʒizə], o participante MB3S recorreu à palavra "dizer", por isso escreveu "diza", com <z>.

Além disso, outros dois participantes, o CS6A (6º ano) e MD3S (3ª série EM), recorreram a uma mesma palavra, "tesão", para escrever o estímulo ['tezə]. No entanto, o raciocínio que eles utilizaram seguiram caminhos diferentes. O aluno CS6A escreveu "tesa", com <s>, por lembrar de "tesão" e achar esta palavra errada, no sentido de inapropriada. Por outro lado, o aluno MD3S escreveu "teza", com <z>, também por lembrar de "tesão", mas argumentou que, por ser uma palavra inapropriada e ele não queria desrespeitar a pesquisadora, "teza". Além dos aspectos linguísticos envolvidos no processo de escrita ortográfica, há também a singularidade do aluno e relação específica que ele desenvolve com determinados itens lexicais.

Na análise quantitativa, notamos que os participantes da 3ª série do EM optaram por escrever <z>, de menor frequência de tipo, e não o <s>, de maior frequência. Isso contrariou um dos pontos da hipótese desta tese, pois esperava-se que, quanto maior a frequência de um grafema, maior a preferência dos estudantes na escrita da irregularidade ortográfica. Ao observar as respostas dos alunos deste ano escolar em relação à escrita de pseudopalavras com o fonema /z/, podemos observar que os participantes apresentaram uma tendência a relacionar o <s> ao fonema de /s/ e apenas o <z> ao fonema /z/, sem considerarem o contexto intervocálico. Nas respostas dos participantes EP3S, KC3S, VB3S, YS3S, GC3S e MD3S relacionadas aos quatro estímulos, observamos uma tendência à escrita de <z>, e não de <s>, pois este, a partir da percepção deles, representa o fonema /s/. Há respostas como: '"Diza", porque essa puxa mais para a parte do 'z', de "za", não puxa muito para a parte do 's'' (participante GC3S); "Por causa da expressão 'Zê', com "s" seria "guissa" [se referindo ao som [s]], com 'z' é 'guiza'" (participante MD3S), por exemplo. Esta ação de dar voz aos sujeitos da pesquisa, além de trazer interessantes reflexões dos caminhos que podem ser percorridos no momento da escrita ortográfica, auxilia, também, na compreensão dos dados analisados quantitativamente na seção anterior.

Outro ponto importante que se destacou durante a análise de dados se refere a casos de escritas que fogem do padrão de escrita do português brasileiro, como (1) consoantes duplas, (2) final de palavra com consoantes que, geralmente, não assumem esta posição; (3) uso frequente de <y>; e (4) uso de <h> para representar o fonema /h/. Foram encontrados casos como: (1) "geffo", "sippa", "senni", "gieddi"; (2) "giv", "gied"; (3) "ciby", "seny", "gyepo", "foone"; (4) "horu". Estes casos podem ter sido motivados pelos desenhos de Digimons utilizados para apresentar as pseudopalavras. O *Digimon* é uma franquia japonesa de desenhos lançada no Brasil em 1999. Os nomes dos *Digimons* seguem tendências como estas quatro encontradas nos dados. Portanto, alguns participantes podem ter seguido as características da escrita dos nomes dos *Digimons*, por isso, foram encontrados estes padrões que não são da escrita do PB. Isso é muito interessante pois, mais uma vez, demonstra que o processo de escrita é múltiplo e por conhecimentos em diversas camadas. Futuras pesquisas poderão investigar a influência da língua estrangeira na nomeação de imagens e escrita de palavras.

Nesta seção, observamos diversos caminhos que podem ser seguidos e diferentes conhecimentos utilizados no processo da escrita de não-palavras. As respostas analisadas fazem parte da consciência metalinguística de cada participante, ou seja, integram a capacidade deles de refletirem sobre a própria língua. Observamos que o processo de escrita é múltiplo, pois envolve diferentes conhecimentos e unidades da língua, como os fonemas, os grafemas, os

nomes das letras, a relação direta entre som e letra, o conhecimento ortográfico, a relação à ortografia de outras palavras e a aspectos gráficos das imagens observadas. O ato de escrever não é apenas o ato de grafar letras, mas sim um processo múltiplo que envolve diversos conhecimentos, linguísticos e não linguísticos. Estas observações contribuem para confirmar a proposição da IMP de que a aprendizagem da ortografia é motivada por múltiplos padrões, que podem ser gráficos, fonológicos, ortográficos, lexical (Treiman; Kessler, 2014; Treiman et al, 2018; Castro; Couto, 2021; Oliveira; Castro; Couto, 2023).

Futuros trabalhos poderão ser desenvolvidos com palavras reais e com participantes de outros anos escolares da educação regular, da Educação de Jovens e Adultos (EJA) e também do Ensino Superior. Também seria interessante a análise de outros sons e contextos. Assim, o processo de escrita de palavras reais poderá ser investigado em maior profundidade observando se a consciência metalinguística pode se modificar a depender do nível de escolaridade.

7.2.4 As variações na escrita de palavras desconhecidas

Além disso, outros dados que ressaltam a imaginação e trazem à tona hipóteses de um estudante no momento da escrita de uma palavra desconhecida. Ainda que estas sejam palavras reais, como “ojeriza” e “cerne”, por serem de baixa frequência, muitos alunos, no momento da coleta de dados, perguntavam o que elas significavam. No Quadro 23, a seguir, há uma síntese das principais versões encontradas nos dados.

Quadro 23 - Versões de escrita de “ojeriza” e “cerne”

Palavra	**Versões da palavra**		
Ojeriza	hogeriza	ugeriza	rogeriza
	hogerisa	rugeria	orgerisa
	nogeriza	oxeriza	ogerisa
Cerne	cerve	cermie	serme
	sernie	ferme	ceme

Fonte: *Corpus* da pesquisa.

Diante de termos desconhecidos, os alunos recorreram a padrões possíveis na escrita, como o uso de <h> no início de “ojeriza”. Além disso, a escrita “cermie” e “sernie” pode ter relação com a palavra real “série”. Por fim, a escrita *oxeriza, com <x>, pode ter relação com a percepção do participante, ou seja, ele pode ter escutado o som de [ʃ], e não de [ʒ]. Esses sons

são semelhantes, sendo o vozeamento a única diferença entre eles, de modo que o primeiro é desvozeado e o segundo vozeado. Dessa forma, como o som [ʃ] pode ser representado por <x> e <ch> e, por ter percebido este som, o aluno optou por <x>. Portanto, é evidente como os alunos recorrem a outros padrões da língua para escrever palavras desconhecidas.

7.3 Resumo do Capítulo

Neste Capítulo, analisamos e discutimos os resultados da pesquisa. Buscamos investigar se a escrita ortográfica de grafemas concorrentes irregulares é influenciada por múltiplos fatores, como o (1) ano escolar, a (2) relação fonema-grafema, a (3) frequência de tipo e a (4) frequência de ocorrência. Para isso, na primeira parte, realizamos a análise quantitativa dos dados. Primeiramente, analisamos os dados da pseudopalavras. Os resultados desta primeira análise nos mostraram que a escrita de grafemas concorrentes irregulares pode ser influenciada, de forma combinada, pelo ano escolar, pela relação fonema-grafema e pela frequência de tipo. Identificamos que há uma tendência pela escrita de grafemas com maior frequência de tipo.

Na segunda parte da análise quantitativa, analisamos os dados das palavras reais. Os resultados indicaram que quanto maior a frequência de tipo e maior a frequência de ocorrência, menor é a probabilidade de escrever uma palavra incorretamente, ou seja, sem seguir as regras ortográficas, ao passo que quanto menor a frequência de ocorrência e a frequência de tipo, maior a probabilidade de errar a escrita de grafemas concorrentes irregulares.

Como discussão geral da análise qualitativa, concluímos que é preciso olhar para a frequência, assim como para outros fatores durante o processo de escrita de grafemas concorrentes irregulares. A frequência pode atuar de forma diferente a depender da relação fonema-grafema. Por exemplo, nas palavras escritas com <g> e o <j>, há a tendência de o participante recorrer à escrita de <g> (maior frequência de tipo) independentemente se a palavra é de maior ou menor frequência de ocorrência. Nas palavras com <c> e <s>, o estudante tende a considerar mais a frequência de ocorrência do que a de tipo.

Por fim, na análise qualitativa, observamos, por meio das respostas dos participantes em relação à escrita das pseudopalavras, que o processo de escrita é complexo e envolve muitos níveis e unidades linguísticas. Isso ocorre, pois, de acordo com as respostas dos participantes à pergunta do que os levou a escrever determinada letra, diferentes conhecimentos são ativados, como os fonemas, os grafemas, os nomes das letras, a relação direta entre som e letra, o conhecimento ortográfico, a relação à ortografia de outras palavras, a aspectos gráficos das

imagens observadas e os conhecimentos prévios sobre os desenhos animados utilizados na metodologia.

A partir das discussões desenvolvidas, a hipótese desta pesquisa foi confirmada. Ou seja, confirmamos que a escrita ortográfica de grafemas concorrentes irregulares é influenciada por múltiplos fatores, como o (1) ano escolar, a (2) relação fonema-grafema, a (3) frequência de tipo e a (4) frequência de ocorrência. No Capítulo, a seguir, há as Considerações finais desta tese.

8 CONSIDERAÇÕES FINAIS

O objetivo geral desta tese foi investigar a influência do ano escolar, da relação fonema-grafema e da frequência de tipo e de ocorrência na escrita de grafemas concorrentes irregulares de alunos de uma escola de Belo Horizonte – MG. Para que o objetivo fosse alcançado, estudamos três casos de irregularidades ortográficas, a saber: (1) o fonema /s/ pode ser representado pelos grafemas <c> ou <s> no contexto diante de <i> e <e> em início de palavra, como em "Cibele", “sigilo”, “seco” e “cego”; (2) o fonema /ʒ/ pode ser representado por dois grafemas (<g> e <j>) em um mesmo contexto, quando diante de <i> ou <e>, como em “gigante”, “jiboia”, “geleia” e “jerico”; por fim, (3) o fonema /z/ pode ser representado pelos grafemas <VsV> e <z> em contexto intervocálico, como “natureza”, “mesa”, “azedo” e “frase”.

Para verificar a frequência de tipo dos grafemas <c> e <s>; <g> e <j>; <VsV> e <z> nos contextos descritos, utilizamos o Vocabulário Ortográfico Comum da Língua Portuguesa – VOC– (Bechara, 2017). A partir desta verificação e do Teste Qui-quadrado, classificamos os grafemas quanto à frequência de tipo da seguinte forma: <c> menor frequência de tipo e <s> maior frequência de tipo; <j> menor frequência de tipo e <g> maior frequência de tipo; <VsV> menor frequência de tipo e <z> maior frequência de tipo. Após isso, elaboramos pseudopalavras com os grafemas descritos. As pseudopalavras compuseram a “Tarefa 1 – pseudopalavras".

Após a classificação da frequência de tipo e a elaboração de pseudopalavras, verificamos a frequência de ocorrência de palavras reais em dois corpora, o LexPorBR – Infantil (Estivalet *et al.*, 2023) e o LexPorBR não - infantil (Estivalet *et al.* 2019). Selecionamos palavras com os grafemas descritos anteriormente e classificamos as em maior frequência de ocorrência e menor frequência. As palavras reais compuseram a “Tarefa 2 – palavras reais”.

O arcabouço metodológico desta tese foi pautado na abordagem mista quali-quantitativa (Paiva, 2019), do tipo de pesquisa de campo experimental (Lakatos; Marconi, 2003) e descritiva (Andrade, 2010). Os dados foram coletados em turmas do 3º ano do Ensino Fundamental (EF), no 6º ano do EF, no 9º ano do EF e na 3ª série do Ensino Médio (EM) na escola pública na cidade de Belo Horizonte – MG.

A análise de dados foi realizada em duas etapas, a quantitativa e a qualitativa. A análise quantitativa foi desenvolvida em duas partes, a saber: Parte 1 - Frequência de tipo: pseudopalavras; Parte 2 - Frequência de ocorrência: palavras reais. Na parte 1, analisamos se a frequência de tipo, a relação fonema-grafema e o ano escolar poderiam influenciar na escrita de grafemas concorrentes irregulares estudados nesta tese. Os resultados do modelo linear generalizado misto indicaram que a escrita do grafema com menor frequência de tipo varia a

depender da relação fonema-grafema. Ou seja, não podemos considerar a frequência de tipo isoladamente, é importante levar em consideração a relação fonema-grafema e o ano escolar.

Na parte 2, analisamos se a frequência de tipo, a relação fonema-grafema, o ano escolar e a frequência de ocorrência poderiam influenciar na escrita de grafemas concorrentes irregulares estudados nesta tese. Os resultados indicaram que os alunos, em cada ano escolar, têm a tendência de errar mais em palavras de baixa frequência de ocorrência e também de baixa frequência de tipo.

Na análise qualitativa, damos voz às hipóteses dos alunos em relação à escrita de grafemas concorrentes irregulares. Observamos que o processo de escrita é múltiplo, pois envolve diferentes conhecimentos e níveis da língua, como os fonemas, os grafemas, os nomes das letras, a relação direta entre som e letra, o conhecimento ortográfico, a relação à ortografia de outras palavras, a aspectos gráficos das imagens observadas. Estas observações confirmam a concepção da IMP de que a aprendizagem da ortografia é motivada por múltiplos padrões, que podem ser gráficos, fonológicos, ortográficos, lexical (Treiman; Kessler, 2014; Treiman *et al*, 2018; Castro; Couto, 2021; Oliveira; Castro; Couto, 2023). Portanto, foi perceptível que, além dos aspectos linguísticos envolvidos no processo de escrita ortográfica, há também a singularidade do aluno e relação específica que ele desenvolve com determinados itens lexicais.

Os resultados desta pesquisa, portanto, confirmaram a hipótese de que a escrita ortográfica de grafemas concorrentes irregulares é influenciada por múltiplos fatores, como (a) o ano escolar, (b) a relação fonema-grafema, (c) a frequência de tipo e (d) a frequência de ocorrência. A frequência, de tipo e de ocorrência, tem a sua atuação no processo de escrita, mas ela precisa estar em interação com outros fatores, como o ano escolar e a relação – fonema grafema. Esse resultado é importante para pesquisas futuras, para que não considerem apenas a frequência como variável de impacto no processo de escrita, mas que possam avaliar a relação desta com outros fatores, visto que as motivações de escrita vão além do sistema linguístico.

Compreender o processo de aquisição e produção escrita do aluno é o primeiro passo para melhoria das propostas em sala de aula. Esta pesquisa evidenciou, sobretudo, que a frequência é relevante e, portanto, pode ser considerada nas propostas de ensino. Este resultado traz contribuições para orientações curriculares, avaliações em larga escala, produção e análise de livros didáticos. Outro aspecto é que o aluno estabelece uma relação variável com conhecimentos diversos, como o semântico e visual, por exemplo. Sabendo-se disso, as aulas de Língua Portuguesa na educação básica em relação à ortografia podem ir além de apenas grafar as palavras estudadas, soletrar, repetir e preencher lacunas. O professor de português pode trabalhar com os alunos as relações ortográficas por meio de uma investigação dos padrões

que mais ocorrem na ortografia do português brasileiro. Além disso, estimular os alunos a recorrerem a outras palavras quando surgirem dúvidas na escrita ortográfica de palavras e realizar associações pode colaborar para este processo. Por fim, ressaltamos que a ortografia precisa ser tratada como objeto de investigação, principalmente, nas salas de aula da Educação Básica.

Esta tese abre caminhos para futuras pesquisa, entre as quais se destacam: investigar o impacto das imagens de palavras na escolha lexical realizada para representá-las; avaliar, a exemplo do léxico, as relações entre morfologia e frequência; e analisar padrões fonológicos adicionais em sua relação com a escrita e a escolha de grafemas ao longo dos anos escolares, especialmente em estudos de natureza longitudinal. Ademais, sugere-se investigar o papel da frequência na aprendizagem da ortografia de alunos da Educação de Jovens e Adultos (EJA). Ressalta-se, ainda, a importância de análises linguísticas computacionais em diferentes línguas para uma classificação mais precisa da transparência e opacidade da ortografia. Futuras pesquisas também poderão explorar a influência de línguas estrangeiras na nomeação de imagens e na escrita de palavras.

REFERÊNCIAS

ALTMILLER, R.; TREIMAN, R; KESSLER, B. Double trouble: Using spellings of different lengths to represent vowel length in English. **Journal of Experimental Child Psychology**, 231, p.105-649, 2023

ANDRADE, M. M. de. **Introdução à metodologia do trabalho científico**: elaboração de trabalhos na graduação – 10. ed. – São Paulo: Atlas, 2010.

APEL, K.; HENBEST, V. C.; MASTERSON, J. Orthographic knowledge: clarifications, challenges, and future directions. **Reading and Writing**, v. 32, p. 873-889, 2019.

BATISTA, A. O.; CAPELLINI, S. A. Desempenho ortográfico de escolares do 2º ao 5º ano do ensino privado do município de Londrina. **Psicologia Argumento**, v. 29, n. 67, 2017.

BECHARA, E. (coord.). **Vocabulário Ortográfico da Língua Portuguesa**. 5.ª edição [adaptada ao VOC]. Rio de Janeiro: Academia Brasileira de Letras, 2017.

BONINI, J. B.; KESKE-SOARES, M. Pseudopalavras para Terapia Fonológica: uma nova abordagem terapêutica. **Codas**, 30(6), 2018.

BRASIL, Ministério da Educação e do Desporto. **Parâmetros Curriculares Nacionais (PCN)** Língua Portuguesa / Secretaria de Educação Fundamental. Brasília, 1997.

BRASIL. **Base Nacional Comum Curricular**. Brasília: MEC, 2017

BRASIL. **Base Nacional Comum Curricular**. Brasília: MEC, 2018.

BRASIL. Congresso. Senado. **Constituição (1996)**. Lei nº 9394, de 20 de dezembro de 1996. Estabelece as diretrizes e bases da educação nacional. Lei Nº 9.394, de 20 de Dezembro de 1996. Brasília, 20 dez. 1996. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l9394.htm. Acesso em: 05 dez. 2023.

BRASIL. **Decreto nº 6583, de 29 de setembro de 2008**. Decreto Nº 6.583, de 29 de Setembro de 2008: Promulga o Acordo Ortográfico da Língua Portuguesa, assinado em Lisboa, em 16 de dezembro de 1990. Brasília, 2008.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Sinopse Estatística da Educação Básica 2022**. Brasília: Inep, 2023.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Saeb 2021: Indicador de Nível Socioeconômico do Saeb 2021: nota técnica. Brasília, DF: Inep, 2023.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – Inep. **Consulta ao Índice de Desenvolvimento da Educação Básica**, 2024. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb/resultados. Acesso em: 2 out. 2024.

BYBEE, J. **Língua, uso e cognição**. Tradução: Maria Angélica Furtado da Cunha; Sebastião Carlos Leite Gonçalves. São Paulo: Cortez, 2016.

BYBEE, J. **Phonology and Language Use**. Cambridge: Cambridge University Press, 2001.

CAGLIARI, L. C. **Alfabetização e linguística**. 10. ed. Scipione. São Paulo, 1997.

CAPOVILLA, F. C.; VARANDA, C.; CAPOVILLA, A. G. S.. Teste de Competência de Leitura de Palavras e Pseudopalavras: normatização e validação. **Psic**, São Paulo, v. 7, n. 2, p. 47-59, dez. 2006.

CARAVOLAS *et al*. Different patterns, but equivalent predictors, of growth in reading in consistent and inconsistent orthographies. **Psychological Science**, 24, 1398–1407, 2013.

CARDOSO-MARTINS, Cláudia; CORREA, Marcela Fulanete. O desenvolvimento da escrita nos anos pré-escolares: questões acerca dos estágios silábicos. **Psicologia: Teoria e Prática**, v.24, n.3, p. 279-86, 2008.

CARVALHO, G. T.. Grafema. *In:* FRADE, I. C. A. S; VAL, M. da G. C. G; BREGUNCI, M. das G. C. **Glossário Ceale de termos de Alfabetização, leitura e escrita para educadores**. Belo Horizonte, CEALE/Faculdade de Educação da UFMG, 2014. Disponível em: <http://ceale.fae.ufmg.br/app/webroot/glossarioceale/verbetes/grafema&gt>. Acesso em:02 jun.2019.

CASTRO, M. de. **Ortografia no Ensino Fundamental II:** múltiplos padrões e (re)escrita textual. (Doutorado em Estudos Linguísticos). 2022. 181f. Curso de Pós-Graduação em Estudos Linguísticos da Faculdade de Letras da Universidade de Minas Gerais, Belo Horizonte, 2022.

CASTRO, M.; COUTO, A. L. S. A integração de múltiplos padrões como uma perspectiva teórica diferenciada à aprendizagem e ao ensino de ortografia. *In:* **Anais do I Congresso Nordestino de Linguística Aplicada**. Aracaju: Ed. dos Autores, 2021. p. 1334-1346.

CHETAIL F. Reconsidering the role of orthographic redundancy in visual word recognition. **Front. Psychol**. 6:645, 2015.

CHETAIL, F. What do we do with what we learn? Statistical learning of orthographic regularities impacts written word processing. **Cognition**, 163, 103–120, 2017.

COLTHEART, M *et. al.* DRC: a dual route cascaded model of visual word recognition and reading aloud. **Psychological Review**, 108(1), 204–256, 2001.

COLTHEART, M. Dual route and connectionist models of reading: an overview. **London Review of Education**, v. 4, n.1, p. 5-17, 2006.

CORTEZ A.C.M. **Funções executivas e leitura de palavras e pseudopalavras em crianças alfabéticas** [dissertação]. São Paulo: Faculdade de Medicina, Universidade de São Paulo; 2018.

COSTA, L; SANTOS, D.; CARDOSO, N. Perspectivas sobre a Linguateca. **Linguateca**. 2008.

COUTINHO, I. de L. **Gramática Histórica**. 7. ed. Rio de Janeiro: Ao Livro Técnico. ed. rev. / 1984.

COUTO, A. L. S. **A ortografia nos livros didáticos do 6º e 7º anos do Ensino Fundamental** [manuscrito]. Dissertação (mestrado) – Universidade Federal de Minas Gerais, Faculdade de Letras, 2020.

CRISTÓFARO SILVA, T. **Dicionário de Fonética e Fonologia**. Colaboradoras: Daniela Oliveira Guimarães e Maria Mendes Cantoni. – 1. Ed., 2ª reimpressão. São Paulo: Editora Contexto, 2017.

CRISTÓFARO SILVA, T. **Fonética e Fonologia do Português**: Roteiro de Estudos e Guia de Exercícios. 10 ed. São Paulo: Contexto, 2013.

CRISTÓFARO SILVA, T.; OLIVEIRA-GUIMARAES, D. M. L. A aquisição da linguagem falada e escrita: o papel da consciência linguística. **Letras de Hoje** (Impresso), v. 48, p. 316-323, 2013.

CRISTÓFARO-SILVA, T.; ALMEIDA, L. S., FRAGA, T. ASPA: a Formulação de um Banco de Dados de Referência da Estrutura Sonora do Português Contemporâneo. *In:* **XXV Congresso da Sociedade Brasileira de Computação**, 2005, São Leopoldo. Anais do XXV Congresso da Sociedade Brasileira de Computação (CD-Room). São Leopoldo: Sociedade Brasileira de Computação, v. 1. p. 2268-2277. 2005.

EHRI, L. C. Learning to Read Words: Theory, Findings, and Issues. **Scientific Studies of Reading**, 9(2), 167–188, 2005.

ESTIVALET, G. *et. al*. **Léxico do Português Brasileiro Infantil - LexPorBR-Infantil**. 2023. Disponível em: https://www.lexicodoportugues.com/infantil/. Acesso em: 07 ago. 2023.

ESTIVALET, G. L.. **Léxico do Português Brasileiro - LexPorBR**. 2019. Gustavo Lopez Estivalet. Disponível em: https://www.lexicodoportugues.com/. Acesso em: 08 ago. 2023.

ESTIVALET, G. L.; MEUNIER, Fanny. Corpus psicolinguístico Léxico do Português Brasileiro. **Soletras: Revista do Programa de Pós-Graduação em Letras e Linguística** – PPLIN Faculdade de Formação de Professores / Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro, n. 33, p. 212-229, 30 jun. 2017.

FAY, J; HEIN, K.; GHAYOOMI, M. Wortfrequenz und Rechtschreibleistung. Erschienen in: **Muttersprache**, 2015.

FERREIRA *et al*. **Vocabulário Ortográfico Comum da Língua Portuguesa**. Praia: Instituto Internacional da Língua Portuguesa / Comunidade dos Países de Língua Portuguesa. 2017

FERREIRA, A. A; BUSSE, S. Processos fonológicos e escrita ortográfica em produções textuais do ensino fundamental. **Domínios de Lingu@gem**, Uberlândia, v. 13, n. 1, p. 233–256, 2019.

FERREIRO, E.; TEBEROSKY, A. **Psicogênese da língua escrita**. Porto Alegre: Artes Médicas, [1984] 1999.

FRANKE, T. M.; HO, T.; CHRISTIE, C. A. The Chi-Square Test: Often Used and More Often Misinterpreted. **American Journal of Evaluation**, v. 33, n. 3, p. 448-458, 2012.

GERMANI, M. M. **A escrita no Ensino Superior:** uma análise desenvolvida com alunos do quinto ano do curso de direito. 2017. 210 f. Dissertação (Mestrado) - Curso de Programa de Pós-Graduação em Educação, Faculdade de Ciências e Tecnologia, Universidade Estadual Paulista, Presidente Prudente, 2017.

GODOY, M. C.; NUNES, M. A. **Estatística para as Ciências da Linguagem**. Em preparação.

GRAHAM, S.; SANTANGELO, T. Does spelling instruction make students better spellers, readers, and writers? A meta-analytic review. **Springer Science+Business Media Dordrecht**. Read Writ, (27), p.1703-1743, 2014.

GRIES, S. **Estatística com R para a linguística**: uma introdução prática. Tradução: Melo, Heliana *et al.* Belo Horizonte. FALE/UFMG, 2019.

GUIMARÃES, D.; CRISTÓFARO SILVA, T.; GOMES, C.. Aquisição do plural irregular no Português Brasileiro: uma abordagem baseada em exemplares. **Revista Linguística**. Volume 16. Número Especia Comemorativo. p. 622-645. 2020

GUIMARAES, S. B.; MOTA, M. M.. Consciência morfológica e ortografia. Uma relação para além da consciência fonológica? **Estud. pesqui. psicol**., Rio de Janeiro, v. 18, n. 2, p. 608-623, ago., 2018.

HENBEST, V. S. *et al.* The Relation Between Linguistic Awareness Skills and Spelling in Adults: A Comparison Among Scoring Procedures. **Journal of Speech, Language, and Hearing Research.** Vol. 63, 2020.

HUBACK, A. P. **Efeitos de frequência nas representações mentais**. Tese (Doutorado em Linguística) - Faculdade de Letras da Universidade Federal de Minas Gerais (UFMG). Belo Horizonte. 2007.

IAIA M, MARINELLI CV, VIZZI F, ANGELELLI P. Aquisição ortográfica em uma ortografia consistente: O efeito facilitador da frequência silábica em soletradores novatos. **PLoS UM** 17(11), 2022.

JENSEN, A. R. Spelling errors and the serial-position effect. **Journal of Educational Psychology**, v. 53, n. 3, jun. 1962

LAKATOS, E. M.; MARCONI, M. A. **Fundamentos da metodologia científica**. 5. ed. São Paulo: Atlas, 2003

LEMLE, M. **Guia teórico do Alfabetizador**. São Paulo: Ática, [1994] 2009.

LOPES, L. P. da M. Linguística Aplicada e vida contemporânea: Problematização dos Construtos que têm orientado a pesquisa. *In:* LOPES, L. P. da M., P. **Por uma Linguística Aplicada Indisciplinar**. São Paulo: Parábola Editorial. p. 85-108, 2008.

MAJORANO *et al*. Early Literacy Skills and Later Reading and Writing Performance Across Countries: The Effects of Orthographic Consistency and Preschool Curriculum. **Child Youth Care Forum**, 2021.

MARCONI L, OTT M, PESENTI E, RATTI D, TAVELLA M. Lessico elementare: Dati statistici sull'italiano scritto e letto dai bambini delle elementari [Léxico elementar: dados estatísticos para italiano escrito e lido por crianças do ensino fundamental]. **Bolonha**, Itália: Zanichelli; 1993.

MARQUARDT, Valéria Caimi; BUSSE; Sanimar. Um estudo dos erros ortográficos em produções escritas de alunos do 9º ano do Ensino Fundamental. **Revista Educação e Linguagens**, Campo Mourão, v. 4, n. 6, jan./jun. 2015.

MASSINI-CAGLIARI, G.; CAGLIARI, L. C. **Diante das letras:** a escrita na alfabetização. Campinas: Mercado de Letras, 1999.

MIRANDA, A. R. M. Ortografia reflexões sobre a aquisição e o ensino. *In:* Leffa, W ; Ernst, A. (Org.). **Linguagens:** metodologias de ensino e pesquisa. Pelotas: EDUCAT, 2012.

MIRANDA, A. R. M. Um estudo sobre a natureza dos erros (orto)gráficos produzidos por crianças dos anos iniciais. **Educ. rev.**, Belo Horizonte, v. 36, e221615, 2020.

MIRANDA, A. R. M. Um estudo sobre o erro ortográfico. *In:* HEINING, O. L.; FRONZA; C. de A. (org.). **Diálogos entre linguística e educação**. 1 ed. Blumenau: EDIFURB, v. 1, p. 141 162, 2010.

MIRANDA, A. R. M; PACHALSKI, L.; RICHETTI, L. Os dígrafos do português na escrita de alunos dos anos iniciais do Ensino Fundamental. **Fórum Linguístico**, Florianópolis, v. 20, n. 1, p. 8727-8745, 07 fev. 2023.

MORAIS, A. G. de. **Ortografia:** ensinar e aprender. São Paulo: Ática, 1998.

MORAIS, A. G.; TEBEROSKY, A. Erros e transgressões infantis na ortografia do Português. **Discursos**, n.8, 1994, 15-51.

NASCIMENTO, J. F. do; HENZ, R. R. A ortografia e os níveis de escrita: o erro em textos de sujeitos escolarizados. Confluência. Rio de Janeiro: **Liceu Literário Português**, n. 61, p. 226 248, jul.-dez. 2022.

NEVES, Maria Helena de Moura. O acordo ortográfico da língua portuguesa e a meta de simplificação e unificação. **DELTA**, São Paulo, v. 26, n. 1, p. 87-113, 2010 .

NIGRO, L et al. Implicit Learning of Written Regularities and Its Relation to Literacy Acquisition in a Shallow Orthography. **Journal of Psycholinguistic Research**, 44(5), 571–585, 2014.

NOBILE, G. G.; BARRERA, S. Domingos. Desempenho Ortográfico e Habilidades de Produção Textual em Diferentes Condições de Solicitação. **Psicologia: Teoria e Pesquisa**., v.32, 2016.

NÓBREGA, M. J. **Ortografia**. São Paulo: Melhoramentos, 2013.

NUNES, S. M.; SANTOS, R. A. B. dos; BARBOSA, J. B. Análise de desvios de ortografia na escrita de alunos do sétimo ano do ensino fundamental de Uberaba-MG. **Revista do Sel**l, [S. l.], v. 9, n. 1, p. 88–104, 2020.

OLIVEIRA, A.M. Desempenho dos escolares do Ensino Fundamental II e Ensino médio no processo lexical. **Revista Neuropsicologia Latinoamericana**. Vol. 13, n.1, 2021.

OLIVEIRA, D. M. L. **Percurso de construção da fonologia pela criança: uma abordagem dinâmica**. Tese (Doutorado em Estudos Linguísticos) – Faculdade de Letras da UFMG, Belo Horizonte, 2008.

OLIVEIRA, D. M. L.; COUTO; A L S; LACERDA, L. V. Análise ortográfica em textos do Enem: evidências da integração de múltiplos padrões. **Cadernos de Educação** -Pelotas, n. 67, p. 01-25, 2023.

OLIVEIRA, D. M. L; CASTRO, M. de; COUTO, A L S. **Ortografia:** reflexão e múltiplos padrões no ensino e na aprendizagem da língua portuguesa. Campinas: Pontes Editores, 2023. 113 p.

OLIVEIRA, M. A. Conhecimento linguístico e apropriação do sistema de escrita. Belo. Horizonte: **Ceale/Fae/UFMG**, 2005.

OTAKE, S., TREIMAN, R.; YIN, L. Differentiation of writing and drawing by U.S. two- to five-year-olds. **Cognitive Development**, 43, 119–128, 2017.

OUSHIRO, L. **Introdução à Estatística para Linguistas**, v.1.0.1, 2017.

PACTON, S.; PERRUCHET, P.; FAYOL, M.; CLEEREMANS, A. Implicit learning out of the lab: The case of orthographic regularities. **Journal of Experimental Psychology**: General, 130(3), 401–426, 2001.

PAIVA, V. L. M. **O Manual de pesquisa em estudos linguísticos**. São Paulo: Parábola, 2019.

PAULA, A. C.. **Tratamento dos desvios ortográficos mais recorrentes de alunos do 9º ano do Ensino Fundamental**. 189 f. Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Humanidades, Programa de Pós-graduação Profissional em Letras, Mestrado Profissional em Letras, Fortaleza, 2021.

PINHEIRO, A. M. V. Contagem de freqüência de ocorrência de palavras expostas a crianças na faixa pré-escolar e séries iniciais do 1o grau. [Software]. São Paulo: **Associação Brasileira de Dislexia**, 1996.

PINHEIRO, G. M.; ALUÍSIO, S. M. Corpus Nilc: descrição e análise crítica com vistas ao projeto Lacio-Web. **Série de Relatórios do NILC**, 2003.

POLLO, T. C. **The nature of young children's phonological and nonphonological spellings**. 2008. 190 f. Dissertation (PhD) - Psychology, Department of Psychology, Washington University, St. Louis, 2008.

POLLO. T. C.; TREIMAN, R.; KESSLER B. Uma revisão crítica de três perspectivas sobre o desenvolvimento da escrita. **Estudos de Psicologia** (Campinas). Jul; 32(3), p.449–59, 2015.

PRADO, M. N. L. **Aprender ortografia: análise do impacto da instrução e construção de práticas**. 2023. 202 f. Tese (Doutorado) - Curso de Programa de Pós-Graduação em Estudos Linguísticos, Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2023.

R Core Team. **R:** A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2023.

RIBEIRO, M. R.; MARTINS, R. M. F.. Estudo acerca da representação gráfica do fonema /s/ na escrita de alunos do 3º ano do ensino fundamental. **Caligrama: Revista de Estudos Românicos**, [S.l.], v. 25, n. 2, p. 63-84, set. 2020.

RUSSAK, S.; ZARETSKY, E.. Investigating spelling across typologically diverse orthographies. **Written Language & Literacy** 25:1, 2022.

SAGGIOMO, F. L. **Um estudo acerca dos erros ortográficos de alunos do terceiro e do sexto ano do Ensino Fundamental**. 106 p. 2018. Dissertação (Mestrado em Letras) - Programa de Pós Graduação em Letras, Centro de Letras e Comunicação, Universidade Federal de Pelotas, Pelotas, 2018.

SANTAELLA, L.. **Comunicação e pesquisa**: projetos para mestrado e doutorado. São Paulo: Hacker Editores, 2001.

SANTOS, M. T. M. dos; BEFI-LOPES, D. M.. Análise da ortografia de alunos do 4º ano do Ensino Fundamental a partir de ditado de palavras. **CoDAS**. v. 25, n. 3, p. 256-261, 2013.

SANTOS, P. P. dos; SOARES, E. P. M. Uma análise da escrita nos textos de alunos do Ensino Fundamental. fólio - **Revista de Letras**, [S. l.], v. 12, n. 1, 2020.

SARTORI, A. T.; MENDES, L.; COSTA, B. R. Ensino-aprendizagem de língua portuguesa: a questão da ortografia no Ensino Médio. **Caminhos em Linguística Aplicada**, v. 12, p. 120-139, 2015.

SELLA, P. **Erros de grafia em produções de alunos do Ensino Médio:** análise e reflexões. 2017. 144 f. Dissertação (Mestrado em Estudos da Linguagem) – Universidade Estadual do Oeste do Paraná – Unioeste. Cascavel, 2017.

SEYMOUR, *et al.* Foundation literacy acquisition in European orthographies. **British Journal of Psychology**, 94(2), 143–174. 2003.

SOARES, M. **Alfabetização:** A questão dos métodos. 1. Ed., 2ª reimpressão. São Paulo: Contexto, 2018.

SOUZA, A. de O. A escrita no Ensino Fundamental II: uma análise linguística do erro ortográfico à luz dos modelos baseados no uso. **Linguagem & Ensino**, Pelotas, v. 23, n. 2, p. 321-346, abr.-jun. 2020.

SOUZA, C. A.; BRANDÃO, J. D. P.; MELO, C. R. C. de. Análise das dificuldades ortográficas por. Meio de análise de produção de textos. **Educação In Loco**, v.01, n. 01, jan.-jun. 2020.

SOUZA, S. V. de. **Aquisição de consoantes soantes em crianças do ciclo 1 do Ensino Fundamental**. 100f. Tese (Doutorado) - universidade Estadual Paulista (Unesp), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto, 2019.

STELLA V; JOB R. Frequenza sillabica e frequenza di lemmi della lingua italiana scritta [Valores de frequência silábica e lema para palavras italianas escritas]. **Jornal Italiano de Psicologia**, 2000; 3: 633–642.

TEIS-ADAMANTE, D. T.; BUSSE, S. A natureza e a frequência de erros gráficos nas séries finais do ensino fundamental e do ensino médio. R. **Letras**, Curitiba, v. 24, n. 44, p. 35-56, jan./jun. 2022

TEIS-ADAMANTE, D. T.; PARISE, L. T. G.. Ortografia no livro didático: objeto de conhecimento e ensino. **Mandinga –Revista de Estudos Linguísticos**, Redenção-CE, v. 02, n. 01, p. 111-133, jan./jun. 2018.

TOLEDO, C. V. S. **Relações múltiplas entre oralidade e escrita:** vogais médias e róticos. 2023. 198 f. Tese (Doutorado) - Curso de Programa de Pós-Graduação em Estudos Linguísticos, Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2023.

TREIMAN, R. **Beginning to spell:** A study of first-grade children. New York: Oxford University Press, 1993.

TREIMAN, R. *et al*. Statistical Learning and Spelling: Older Prephonological Spellers Produce More Wordlike Spellings Than Younger Prephonological Spellers. **Child development** vol. 89,4: e431-e443, 2018.

TREIMAN, R. Learning to Spell Words: Findings, Theories, and Issues**. Scientific Studies of Reading**, 21(4), p.265–276, 2017.

TREIMAN, R. Learning to Write Words. **Current Directions in Psychological Science**, 29(5), p.521–526, 2020.

TREIMAN, R. Statistical learning and spelling. **Language, Speech, and Hearing Services in Schools**, v. 49, ago, p. 644-652, 2018.

TREIMAN, R.; CARDOSO-MARTINS, C.; POLLO, T; KESSLER, B. Statistical learning and spelling: Evidence from Brazilian prephonological spellers. **Cognition**, v. 182, p. 1–7, 2019.

TREIMAN, R; DECKER, K; KESSLER, B. Adults' sensitivity to graphotactic differences within the English vocabulary. **Applied Psycholinguistics**, 1–16, 2018.

TREIMAN, R; KESSLER, B. **How children learn to write words**. New York, NY: Oxford University Press. 2014.

TREIMAN, R; KESSLER, B. Statistical Learning in Word Reading and Spelling across Languages and Writing Systems. **Scientific Studies of Reading**, 26:2, 139-149, 2022.

WINTER, B. **Statistics for Linguists**: An Introduction Using R (1st ed.). Routledge, 2019.

ZACHARIAS-CAROLINO, A. G.; OSTI, A. Desempenho na escrita de estudantes pertencentes aos anos finais do ensino fundamental I. **Rev. psicopedag**., São Paulo, v. 37, n. 114, p. 314-326, dez.2020.

APÊNDICES

Apêndice A - Termo de Consentimento Livre e Esclarecido (TCLE) aos responsáveis dos alunos

Prezado(a) responsável,_____, o(a) aluno(a) de sua responsabilidade está sendo convidado(a) como voluntário(a) a participar do projeto de pesquisa **"Ensino de Língua Portuguesa pós-ensino remoto: aquisição da escrita, consciência linguística e ortografia"**, cujo **objetivo geral** é "refletir sobre a aquisição da escrita, ortografia e reflexão linguística nos dados dos alunos, discutindo a partir desses dados o impacto da pandemia no aprendizado de língua materna"

Pedimos a sua autorização para a coleta, o armazenamento e a utilização de dados de fala e de escrita do(a) educando(a), seu(sua) dependente. A participação do(a) seu(sua) dependente é voluntária e ele não terá despesas para participar e também não receberá remuneração pela participação. Além disso, seu(sua) dependente tem a liberdade de recusa e de desistência em qualquer momento da pesquisa.

Para que o objetivo da pesquisa seja alcançado, adotaremos os seguintes procedimentos: gravação da voz e coleta de escrita. Na coleta de escrita, seu(sua) dependente escreverá um texto dissertativo sobre um tema previamente selecionado pela pesquisadora. Esta coleta será realizada na sala de aula com todos os alunos da turma em um horário previamente selecionado pelo professor e pela direção da escola. Na coleta de dados de fala, seu(sua) dependente será convidado pela pesquisadora a ir, individualmente, para uma sala onde responderá a perguntas sobre a sua escrita ortográfica. As perguntas, realizadas pela pesquisadora, e as respostas do participante, seu(sua) dependente, serão gravadas. Esta coleta será realizada em uma sala com mesa e cadeira para proporcionar conforto durante o tempo de coleta. Os instrumentos utilizados na gravação são: um gravador de voz e um computador. A gravação terá tempo mínimo de 10 minutos e máximo de 20 minutos. Seu(sua) dependente poderá interromper a gravação de voz a qualquer momento, caso sinta necessidade. Após a coleta, os dados de fala e de escrita serão armazenados em um servidor de uso exclusivo da pesquisadora. Em nenhum momento, a identidade do estudante por quem você é responsável será exposta.

Os riscos envolvidos na pesquisa são: (1) quebra de sigilo dos dados de escrita e das gravações; (2) cansaço do participante durante a pesquisa; (3) constrangimento em relação à ortografia das palavras.

Para sanar tais riscos, serão feitos todos os esforços possíveis, tais como: (1) serão atribuídos códigos no lugar do participante, com intuito de manter o anonimato, além do mais, os textos coletados serão armazenados em armário próprio na Faculdade de Letras da UFMG, durante 5 a 10 anos; (2) o participante terá o livre arbítrio de desistir, a qualquer momento, da pesquisa, ou de pedir uma pausa durante as atividades; (3) não haverá nenhum julgamento em relação à ortografia ou à fala do participante. Em caso de danos provenientes da pesquisa, você poderá buscar indenização nos termos da Resolução 466/2012 do Conselho Nacional de Saúde.

Este projeto contribuirá para a construção de corpus linguísticos com dados de escrita, de alunos matriculados em diversos segmentos educacionais, como o Ensino Fundamental I e II, Ensino Médio. Além disso, os dados obtidos poderão contribuir para a compreensão da trajetória de aprendizagem da ortografia do português brasileiro no contexto pós-ensino remoto. Esta pesquisa também auxiliará na identificação dos problemas relacionados às dificuldades na aprendizagem da língua portuguesa e, a partir disso, direcionar a possíveis atividades que poderão auxiliar os alunos na escrita ortográfica. Por fim, pesquisas futuras podem ser desenvolvidas a partir dos dados coletados nesta pesquisa.

Este termo de consentimento terá duas vias originais, sendo que uma será arquivada pelos pesquisadores responsáveis e a outra será fornecida a você. Os pesquisadores tratarão a sua identidade com padrões profissionais de sigilo, atendendo à legislação brasileira (Resolução 466/2012, do Conselho Nacional de Saúde, Resolução 510/2016 e suas complementares), utilizando as informações somente para fins acadêmicos e científicos.

Eu, _____, portador (a) do documento de Identidade _____, responsável pelo (a) aluno (a) _____, matriculado na escola _____, no ano escolar _____, fui informado (a) dos objetivos, métodos, riscos e benefícios da projeto intitulado de **"Ensino de Língua Portuguesa pós-ensino remoto: aquisição da escrita, consciência linguística e ortografia**", de maneira clara e detalhada, e esclareci minhas dúvidas. Sei que a qualquer momento poderei solicitar novas informações e modificar minha decisão de autorizar a participação do (a) educando (a), por quem sou responsável, se assim o desejar.

☐ Concordo que os **dados de voz** do aluno, por quem sou responsável, sejam utilizados nesta pesquisa.

☐ Concordo que os **dados de escrita** do aluno, por quem sou responsável, sejam utilizados nesta pesquisa.

☐ Concordo que **os dados do aluno**, por quem sou responsável, possam ser utilizados em outras pesquisas, mas serei comunicado pelo pesquisador novamente e assinarei outro termo de consentimento livre e esclarecido que explique para que será utilizado o material.

Rubrica da pesquisadora responsável: _____

Rubrica do (a) responsável pelo (a) aluno (a) menor: _____

Declaro que concordo e autorizo a participação do (a) estudante, por quem sou responsável, nesta pesquisa. Recebi uma via original deste Termo de Consentimento Livre e Esclarecido assinado por mim e pelos pesquisadores, que me deram a oportunidade de ler e esclarecer todas as minhas dúvidas.

Belo Horizonte/MG, _____ de _____ de 20___.

_____
Nome completo do responsável pelo (a) aluno (a) menor

_____
Assinatura do responsável pelo (a) aluno (a) menor

Para dúvidas gerais sobre a pesquisa, você poderá entrar em contato com a pesquisadora responsável:

Nome completo da Pesquisadora Responsável: Daniela Mara Lima Oliveira Guimarães
Endereço: Av. Presidente Antônio Carlos, 6627, Bairro: Pampulha. CEP: 31270-901 /Belo Horizonte – MG
Telefones: (31) 3409 – 6037 E-mail: danielamlog@letras.ufmg.br

_____ _____
Local e data Assinatura da pesquisadora responsável

Em caso de dúvidas, com respeito aos aspectos éticos desta pesquisa, você poderá entrar em contato com o CEP- UFMG:

CEP-UFMG - Comissão de Ética em Pesquisa da UFMG
Av. Antônio Carlos, 6627. Unidade Administrativa II - 2º andar - Sala 2005. Campus Pampulha.
Belo Horizonte, MG – Brasil. CEP: 31270-901. E-mail: coep@prpq.ufmg.br. Tel: 34094592

Apêndice B - Termo de Consentimento Livre e Esclarecido (TCLE) aos alunos maiores de idade

Prezado(a)_____,você está sendo convidado(a) como voluntário(a) a participar do projeto de pesquisa **"Ensino de Língua Portuguesa pós-ensino remoto: aquisição da escrita, consciência linguística e ortografia"**, cujo **objetivo geral** é "refletir sobre a aquisição da escrita, ortografia e reflexão linguística nos dados dos alunos, discutindo a partir desses dados o impacto da pandemia no aprendizado de língua materna". Pedimos a sua autorização para a coleta, o armazenamento e a utilização de dados de fala e de escrita que serão coletados neste projeto. Sua participação na pesquisa é voluntária e você não terá despesas para participar e também não receberá remuneração pela participação. Além disso, você tem a liberdade de recusa e de desistência em qualquer momento da pesquisa.

Para que o objetivo da pesquisa seja alcançado, adotaremos os seguintes procedimentos: gravação da voz e coleta de escrita. Na coleta de escrita, você escreverá um texto dissertativo sobre um tema previamente selecionado pela pesquisadora. Esta coleta será realizada na sala de aula com todos os alunos da turma em um horário previamente selecionado pelo professor e pela direção da escola. Na coleta de dados de fala, você será convidado pela pesquisadora a ir, individualmente, para uma sala onde responderá a perguntas sobre a sua escrita ortográfica. As perguntas, realizadas pela pesquisadora, e as respostas do participante serão gravadas. Esta coleta será realizada em uma sala com mesa e cadeira para proporcionar conforto durante o tempo de coleta. Os instrumentos utilizados na gravação são: um gravador de voz e um computador. A gravação terá tempo mínimo de 10 minutos e máximo de 20 minutos. Você poderá interromper a gravação de voz a qualquer momento, caso sinta necessidade. Após a coleta, os dados de fala e de escrita serão armazenados em um servidor de uso exclusivo da pesquisadora. Em nenhum momento, a sua identidade pessoal será exposta.

Os riscos envolvidos na pesquisa são: (1) quebra de sigilo dos dados de escrita e das gravações; (2) cansaço do participante durante a pesquisa; (3) constrangimento em relação à ortografia das palavras.

Para sanar tais riscos, serão feitos todos os esforços possíveis, tais como: (1) serão atribuídos códigos no lugar do seu nome, com intuito de manter o anonimato, além do mais, os textos coletados serão armazenados em armário próprio na Faculdade de Letras da UFMG, durante 5 a 10 anos; (2) você terá o livre arbítrio de desistir, a qualquer momento, da pesquisa, ou de pedir uma pausa durante as atividades; (3) não haverá nenhum julgamento em relação à sua ortografia ou à sua fala. Em caso de danos provenientes da pesquisa, você poderá buscar indenização nos termos da Resolução 466/2012 do Conselho Nacional de Saúde.

Este projeto contribuirá para a construção de corpus linguísticos com dados de escrita, de alunos matriculados em diversos segmentos educacionais, como o Ensino Fundamental I e II, Ensino Médio. Além disso, os dados obtidos poderão contribuir para a compreensão da trajetória de aprendizagem da ortografia do português brasileiro no contexto pós-ensino remoto. Esta pesquisa também auxiliará na identificação dos problemas relacionados às dificuldades na aprendizagem da língua portuguesa e, a partir disso, direcionar a possíveis atividades que poderão auxiliar os alunos na escrita ortográfica. Por fim, pesquisas futuras podem ser desenvolvidas a partir dos dados coletados neste projeto.

Este termo de consentimento terá duas vias originais, sendo que uma será arquivada pelos pesquisadores responsáveis e a outra será fornecida a você. Os pesquisadores tratarão a sua identidade com padrões profissionais de sigilo, atendendo à legislação brasileira (Resolução 466/2012, do Conselho Nacional de Saúde, Resolução 510/2016 e suas complementares), utilizando as informações somente para fins acadêmicos e científicos.

Eu (seu nome), _____, portador (a) do documento de Identidade _____, matriculado na escola _____, no _____ ano escolar, fui informado (a) dos objetivos, métodos, riscos e benefícios da pesquisa intitulada de "Ensino de Língua Portuguesa pós-ensino remoto: aquisição da escrita, consciência linguística e ortografia", de maneira clara e detalhada, e esclareci minhas dúvidas. Sei que a qualquer momento poderei solicitar novas informações e modificar minha decisão de participação da pesquisa se assim o desejar.

☐ Concordo que os **meus dados de escrita** sejam utilizados nesta pesquisa.

☐ Concordo que os **dados da minha fala** sejam utilizados nesta pesquisa.

☐ Concordo que que os **meus dados** possam ser utilizados em outras pesquisas, mas serei comunicado pelo pesquisador novamente e assinarei outro termo de consentimento livre e esclarecido que explique para que será utilizado o material.

Rubrica da pesquisadora responsável: _____

*Rubrica do (a) participante:*_____

Declaro que concordo e autorizo a minha participação neste projeto. Recebi uma via original deste Termo de Consentimento Livre e Esclarecido assinado por mim e pelos pesquisadores, que me deram a oportunidade de ler e esclarecer todas as minhas dúvidas.

Belo Horizonte/MG,_____ de _____ de 20___.

_____

Assinatura do (a) participante

Para dúvidas gerais ou sobre a pesquisa, você poderá entrar em contato com a pesquisadora responsável:

Nome completo da Pesquisadora Responsável: Daniela Mara Lima Oliveira Guimarães

Endereço: Av. Presidente Antônio Carlos, 6627, Bairro: Pampulha. CEP: 31270-901

Belo Horizonte – MG

Telefones: (31) 3409 – 6037 E-mail: danielamlog@letras.ufmg.br

_____

Assinatura da pesquisadora responsável

Em caso de dúvidas, com respeito aos aspectos éticos desta pesquisa, você poderá entrar em contato com o CEP-UFMG:

CEP-UFMG - Comissão de Ética em Pesquisa da UFMG

Av. Antônio Carlos, 6627. Unidade Administrativa II - 2º andar - Sala 2005. Campus Pampulha.

Belo Horizonte, MG – Brasil. CEP: 31270-901. E-mail: coep@prpq.ufmg.br. Tel: 34094592

Apêndice C - Termo de Assentimento Livre e Esclarecido (TALE)

Prezado (a) aluno (a):_____ do ano escolar____ da Escola _____, você está sendo convidado (a) a participar como voluntário (a) do projeto de pesquisa **"Ensino de Língua Portuguesa pós-ensino remoto: aquisição da escrita, consciência linguística e ortografia"**. Queremos saber sobre como as pessoas estão escrevendo e pensando sobre a língua portuguesa depois do ensino remoto. Precisamos de dados de escrita e dados orais de alunos do 1° ao 9° do Ensino Fundamental e do 1º ao 3º do Ensino Médio. Se você concordar em participar da pesquisa, sua voz será gravada e você deverá escrever um texto e/ou algumas palavras. Seu nome e seus dados nunca serão expostos. Sua participação na pesquisa é voluntária e você não terá despesas e também não receberá remuneração pela participação. Além disso, você tem a liberdade de recusa e de desistência em qualquer momento da pesquisa.

Objetivo: A pesquisa refletirá sobre a aquisição da escrita, ortografia e reflexão linguística nos dados dos alunos, discutindo a partir desses dados o impacto da pandemia no aprendizado de língua materna. Os resultados ajudarão os pesquisadores e professores a entender sobre como as pessoas estão escrevendo e como estão refletindo sobre a língua portuguesa. Sendo assim, você, aluno(a), têm um papel importante no desenvolvimento desta pesquisa.

Os riscos envolvidos na pesquisa são: (1) possível cansaço ao participar das atividades; (2) possível dúvida em relação à escrita ou à pronúncia de algumas palavras. Caso você sinta cansaço ou desconforto, você pode pedir para encerrar ou pausar as atividades. Não se preocupe com a escrita ou com a fala correta das palavras. Você pode escrever e falar livremente, sem se preocupar com erros ou acertos. Seus responsáveis estão cientes dos riscos e benefícios da pesquisa e autorizaram sua participação. Em caso de danos provenientes da pesquisa, você e seus responsáveis poderão buscar indenização nos termos da Resolução 466/2012 do Conselho Nacional de Saúde.

☐ Aceito participar das atividades da pesquisa.

Rubrica da pesquisadora responsável: _____

*Rubrica do (a) participante:*_____

_____

Local e data

_____

Assinatura do (a) aluno (a) menor de idade

Para dúvidas gerais ou sobre a pesquisa, você poderá entrar em contato com a pesquisadora responsável:

Nome completo da Pesquisadora Responsável: Daniela Mara Lima Oliveira Guimarães
Endereço: Av. Presidente Antônio Carlos, 6627, Bairro: Pampulha. CEP: 31270-901 /Belo Horizonte – MG
Telefones: (31) 3409 – 6037
E-mail: danielamlog@letras.ufmg.br

_____ _____

Local e data Assinatura da pesquisadora responsável

Em caso de dúvidas, com respeito aos aspectos éticos desta pesquisa, você poderá entrar em contato com o CEP-UFMG:

CEP-UFMG - Comissão de Ética em Pesquisa da UFMG
Av. Antônio Carlos, 6627. Unidade Administrativa II - 2º andar - Sala 2005. Campus Pampulha.
Belo Horizonte, MG – Brasil. CEP: 31270-901. E-mail: coep@prpq.ufmg.br. Tel: 34094592

Apêndice D – Pseudoplavras do Estudo Piloto

Pseudopalavras	Fonema						Distratores
	/s/		/ʃ/		/ʒ/		
	[siˈpapʊ]	[seˈpapʊ]	[ʃiˈkopə]	[ʃeˈkopə]	[ʒiˈfokə]	[ʒeˈfokə]	[haˈpakʊ]
	[siˈkatʊ]	[seˈkatʊ]	[ʃiˈfatə]	[ʃeˈfatə]	[ʒiˈtolə]	[ʒeˈtolə]	[daˈgotʊ]
	[siˈpalʊ]	[seˈpalʊ]	[ʃiˈfakʊ]	[ʃeˈfakʊ]	[ʒiˈbafʊ]	[ʒeˈbafʊ]	[pɛˈgafʊ]
	[siˈtafə]	[seˈtafə]	[ʃiˈpabə]	[ʃeˈpabə]	[ʒiˈpobʊ]	[ʒeˈpobʊ]	
	[siˈlopə]	[seˈlopə]	[ʃiˈbapʊ]	[ʃeˈbapʊ]	[ʒiˈkapʊ]	[ʒeˈkapʊ]	
Subtotal	**10**		**10**		**10**		**3**
Total	**33**						

Apêndice E - Carta de Anuência do coordenador da instituição

Eu,_____, coordenador(a) pedagógico(a) da _____, autorizo a realização, nesta instituição, da projeto de pesquisa intitulado **"Ensino de Língua Portuguesa pós-ensino remoto: aquisição da escrita, consciência linguística e ortografia"** sob responsabilidade da pesquisadora Prof.ª Dra. Daniela Mara Lima Oliveira Guimarães.

Ciente dos objetivos e da metodologia da pesquisa acima citada, concedo a anuência para o seu desenvolvimento, desde que me sejam assegurados os requisitos abaixo:

- O cumprimento das determinações éticas da resolução 466/12 do CNS.
- A garantia de solicitar e receber esclarecimentos antes, durante e depois do desenvolvimento da pesquisa.

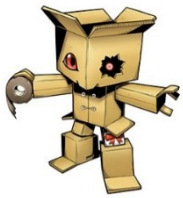
Não haverá nenhuma despesa para esta instituição que seja decorrente da participação nesta pesquisa. No caso de não cumprimento dos itens acima, a instituição tem a liberdade de retirar a anuência da pesquisa a qualquer momento sem penalização alguma.

Belo Horizonte/MG, _____ de _____ de 20____.

_____

Assinatura do coordenador pedagógico

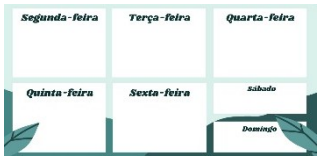
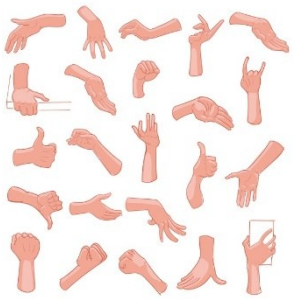
Apêndice F – Pseudopalavras e suas imagens

Fonema	Tipo	Pseudopalavras		
/s/	Palavra-alvo	 ['silʊ]	 ['sikə]	 ['sibɪ]
		 ['sevʊ]	 ['sehɪ]	 '['sipə]
		 ['senɪ]	 ['setʊ]	-
/ʒ/	Palavra-alvo	 ['ʒefʊ]	 ['ʒimə]	 ['ʒepʊ]
		 ['ʒidʃɪ]	 ['ʒikʊ]	 ['ʒivɪ]

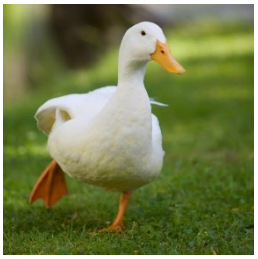
		 ['ʒetʊ]	 ['ʒedʊ]	-
/z/	Palavra-alvo	 ['bezə]	 ['kizə]	 ['nizə]
		 ['hizə]	 ['tezə]	 ['dʒizə]
		 ['kezə]	 ['fezə]	-
Distradores		 ['tʃimə]	 ['kelʊ]	 ['funɪ]

	 ['hoɾʊ]	 ['nekə]	 ['napɪ]
	 ['lumɪ]	 ['kevʊ]	-

Apêndice G – Palavras reais e suas imagens

Fonema	Tipo	Palavras reais		
/s/	Palavra-alvo	 cidade	 cicuta	 certo
		 cerne	 semana	 sequela
		 sistema	 singelo	-
/ʒ/	Palavra-alvo	 general	 gestual	 gigante
		 gincana	 sujeito	 dejeto

		 jipe	 jiló	-
/z/	**Palavra-alvo**	 surpresa	 turquesa	 beleza
		 leveza	 amizade	 ojeriza
		 camisa	 divisa	-

Distradores	 fava	 pato	 pena
	 vaca	 carro	 bola
	 milho	 fogo	-

Apêndice H – Orientações para a aplicação do experimento

- **Orientações pré-coleta:**
  - Você encontrará em cada uma das pastas "Pseudopalavras" e "Palavras Reais" 20 arquivos numerados de **1 a 20**.
  - Para cada aluno deverá ser aberto um arquivo diferente de cada uma das duas pastas.
  - Os arquivos deverão ser abertos em SEQUÊNCIA e em ordem CRESCENTE.
  - ESTAR ATENTO PARA NÃO PULAR NENHUM ARQUIVO.
  - ABRIR os arquivos (um da pasta Pseudopalavras e um da pasta Palavras Reais) e ANOTAR no campo **QUESTIONÁRIO** da folha que será entregue para o aluno O NÚMERO DO ARQUIVO (1 a 20) que foi aberto.
  - ATENÇÃO: você deverá anotar o número para ambas as tarefas: Tarefa 1 - pseudopalavras e Tarefa 2 – palavras reais.

- **Orientações ao recepcionar o aluno:**
  - Dizer: "*Você está participando de uma pesquisa muito importante da UFMG sobre o que as pessoas pensam ao escrever algumas palavras*. *Por isso, fique tranquilo! Não há certo ou errado nesta pesquisa. Queremos entender como você PENSA ao escrever. Então, fique à vontade*".
  - *Esta pesquisa é composta por duas partes. Nas duas partes você ouvirá palavras inventadas ou reais e terá que repeti-las e escrevê-las.*

- **Orientações para o aplicador:**
  - Após escrever o número dos arquivos abertos, ENTREGAR a folha para o aluno e perguntar se ele quer responder a pesquisa a lápis ou à caneta.
  - PEDIR ao aluno para escrever O NOME, a IDADE e a SÉRIE ESCOLAR nos campos reservados
  - ATENÇÃO: Em nenhum momento, você poderá ajudar o aluno na escrita. Evite comentários, sons ou expressões faciais sugestivas de aprovação ou reprovação durante a aplicação do teste. Este momento é do aluno, para ele mostrar o que conhece sobre a ortografia.

- **Orientações para a coleta**

Parte 1 – Tarefa 1 - Pseudopalavras

- Inicie o processo de gravação
- Com o arquivo da pasta Tarefa 1_Pseudopalavras aberto, leia, com o aluno, as instruções do primeiro slide.
- Dizer: *Essa parte da pesquisa será gravada. Então, após ouvir a palavra, você deverá repeti-la em voz alta e escrevê-la na sua folha*.
- Fazer o treinamento e orientar o aluno a escrever a palavra no campo **TREINAMENTO**
- Começar o teste.
- ATENÇÃO: Ficar atento as imagens sublinhadas → você deverá perguntar ao aluno, antes de passar para a próxima palavra, por qual motivo ele escolheu aquelas letras para a escrita da palavra. Caso o aluno não entenda, auxiliar o aluno perguntando, por exemplo, o motivo pelo qual ele optou pela letra S ao invés de C, J ao invés de G, etc.
- NA PALAVRA 15, se o cansaço do aluno for perceptível, pergunte se ele precisa dar uma pausa.
- Ao FINALIZAR a Tarefa 1_Pseudopalavras, feche o respectivo arquivo e SALVE a gravação com o nome do aluno.

- Pergunte ao aluno se ele precisa ir ao banheiro ou beber água antes do próximo teste.

Parte 2- Experimento palavras reais

- ESSA PARTE NÃO DEVERÁ SER GRAVADA
- Com o arquivo da pasta Tarefa 2_PalavrasReais aberto, leia, com o aluno, as instruções do primeiro slide.
- Faça o treinamento e oriente o aluno a escrever a palavra no campo **TREINAMENTO**
- Iniciar o teste.
- Na palavra 15, se o cansaço do aluno for perceptível, perguntar se ele quer dar uma pausa.

➢ **Orientações pós-coleta**

- Ao finalizar o teste, agradeça ao aluno pela participação e comunique que a caneta/lápis com a qual ele realizou o experimento é um presente.
- Pedir para o aluno voltar à sala de aula e que ele convide outro aluno para se dirigir à sala do experimento.
- Guarde a folha utilizada e se prepare para o próximo aluno.

Apêndice I – Folha de resposta

QUESTIONÁRIO:_____

Nome:_____

Idade:_____ Ano/Série escolar:_____

Experimento 1 – pseudopalavras

Treinamento:_____

1- _____
2- _____
3- _____
4- _____
5- _____
6- _____
7- _____
8- _____
9- _____
10- _____
11- _____
12- _____
13- _____
14- _____
15- _____
16- _____
17- _____
18- _____
19- _____
20- _____
21- _____
22- _____
23- _____
24- _____
25- _____
26- _____
27- _____
28- _____
29- _____
30- _____
31- _____
32- _____

QUESTIONÁRIO:_____

Experimento 2 – Palavras reais

Treinamento:_____

1- _____
2- _____
3- _____
4- _____
5- _____
6- _____
7- _____
8- _____
9- _____
10- _____
11- _____
12- _____
13- _____
14- _____
15- _____
16- _____
17- _____
18- _____
19- _____
20- _____
21- _____
22- _____
23- _____
24- _____
25- _____
26- _____
27- _____
28- _____
29- _____
30- _____
31- _____
32- _____

Muito obrigada pela sua participação!

Apêndice J – Scripts das análises estatísticas no R

##
################### **TESE** ##############
##

```
#########Instalar pacotes ##########
install.packages("tidyverse")
install.packages("broom")
install.packages("emmeans")
install.packages("car")
install.packages("ggfortify")
install.packages("lme4")
install.packages("forcats")
install.packages("lmerTest")
install.packages("patchwork")
install.packages("effects")
install.packages("ggrepel")
install.packages("ggpubr")
install.packages("rms")
install.packages("lme4")
install.packages ("Matrix")


########Carregar pacotes###########
library("tidyverse")
library("broom")
library("ggfortify")
library("emmeans")
library("car")
library("ggfortify")
library("lme4")
library("forcats")
library("lmerTest")
library("patchwork")
library("effects")
library("ggrepel")
library("ggpubr")
library("rms")
library("Matrix")
```

```
### Importar dados####
#diretorio dos dados
setwd("C:/Users/annal/Desktop/Doutorado/3-DADOS")
dados.brutos <- read.csv2("dados_analises.csv")

#### Calcular Qui-quadrado no Grupo A e B da frequência de tipo dos grafemas relacionados a cada fonema

##### Fonema /s/ e os grafemas <s> e <c>
ftipo.sa <- c(3682, 9348,5666, 9348)
## Calcular Qui-quadrado
chisq.test(ftipo.sa)


#### Fonema /ja/ e os grafemas <g> e <j>
ftipo.ja <- c(1476, 1920, 444, 1920)
# Calcular Qui-quadrado
chisq.test(ftipo.ja)


#### Fonema /z/ e os grafemas <s> e <z>
ftipo.za <- c(110597, 202630, 92033, 202630)
## Calcular Qui-quadrado
chisq.test(ftipo.za)



#### Calcular Qui-quadrado no Grupo A e B da frequência de ocorrência
## Corpus Infantil e nao infantil

#### Palavras com fonema /sa/
### Corpus Infantil
focorrencia.sa <- c(41918, 312551, 46151, 10879, 63, 38, 49,26)

#### Calcular Qui-quadrado do conjunto de palavras com /sa/
#### Corpus Infantil
chisq.test(focorrencia.sa)

#### Palavras com fonema /ja/ - Corpus Infantil
```

```
focorrencia.ja <- c(8051, 5847, 3667, 552, 19, 57, 13,1)

#### calculando Qui-quadrado do conjunto de palavras com /ja/
#### Corpus Infantil
chisq.test(focorrencia.ja)

#### Palavras com fonema /za/
#### Corpus Infantil
focorrencia.za <- c(17953, 16301, 5813, 8610, 123, 107, 2,72)

#### calculando Qui-quadrado do conjunto de palavras com /ja/
#### Corpus Infantil
chisq.test(focorrencia.za)

#### Palavras com fonema /sa/ - Corpus Nao infantil
focorrencia.nsa <- c(16093, 4668, 14927, 10723, 18, 76, 21,40)

#### calculando Qui-quadrado do conjunto de palavras com /sa/
##### Corpus Nao infantil
chisq.test(focorrencia.nsa)

#### Palavras com fonema /ja/ - Corpus Nao infantil
focorrencia.nja <- c(1849, 206, 1127, 114, 33, 26, 5,1)

#### calculando Qui-quadrado do conjunto de palavras com /ja/
#### Corpus Nao infantil
chisq.test(focorrencia.nja)

#### Palavras com fonema /za/ - Corpus Nao infantil
focorrencia.nza <- c(1220, 1162, 608, 1207, 14, 115, 25,170)

#### calculando Qui-quadrado do conjunto de palavras com /ja/
#### Corpus Nao infantil
chisq.test(focorrencia.nza)


                ############################################################
              ############### INSPECIONAR DADOS ##########################
                ############################################################
```

```
## Nomes das colunas
names (dados.brutos)

## Colunas dos dados
head (dados.brutos)

## Tipo dos dados nas colunas
str(dados.brutos)


###### Mudar variáveis para codificarem como fator #####

## variável participante
dados.brutos$participante <- as.factor(dados.brutos$participante)

## variável ano escolar
dados.brutos$ano <- as.factor(dados.brutos$ano)

## variável estimulo
dados.brutos$estimulo <- as.factor(dados.brutos$estimulo)

## variável frequência de ocorrência - qualitativa categórica
dados.brutos$focorrenciacat <- as.factor(dados.brutos$focorrenciacat)

## variável fonema
dados.brutos$fonema <- as.factor(dados.brutos$fonema)

## variável grafema
dados.brutos$grafema <- as.factor(dados.brutos$grafema)

## variável ftipoescritacat = frequência de tipo categórica (grafema)
dados.brutos$ftipogracat <- as.factor(dados.brutos$ftipogracat)

## variável tipo - pseudo ou palavra real
dados.brutos$tipo <- as.factor(dados.brutos$tipo)

## variável escrita
dados.brutos$escrita <- as.factor(dados.brutos$escrita)

## variável letra = escrita do participante
```

```
dados.brutos$letra <- as.factor(dados.brutos$letra)

## variável ftipoescritacat = frequência de tipo categórica a partir da escrita
dados.brutos$ftipoescritacat <- as.factor(dados.brutos$ftipoescritacat)

## variável certoerrado = erro e acerto da palavra real
dados.brutos$certoerrado <- as.factor(dados.brutos$certoerrado)


## Conferir o tipo das variáveis
str(dados.brutos)

########## Observar os niveis das variaveis ##########
levels(dados.brutos$participante) ## participante
levels(dados.brutos$ano) ## ano escolar
levels(dados.brutos$estimulo) ## estimulo
levels(dados.brutos$focorrenciacat) ## frequência de ocorrência
levels(dados.brutos$fonema) ## fonema
levels(dados.brutos$grafema) ## grafema
levels(dados.brutos$ftipogracat) ## ftipografema categórica
levels(dados.brutos$tipo) ## tipo
levels(dados.brutos$escrita) ## escrita
levels(dados.brutos$letra) ## letra
levels(dados.brutos$ftipoescritacat) ## ftipoescrita categórica
levels(dados.brutos$certoerrado) ## certoerrado

### Observar dados gerais por ano escolar ###
with(dados.brutos,table(ano))

### Observar dados gerais por fonema ###
with(dados.brutos,table(fonema))

### Observar dados gerais por ano escolar e fonema ###
with(dados.brutos,table(fonema, ano))


                ######################################################################
               ################## ESTATISTICA DESCRITIVA, TABELAS e ###############
                ########################## VISUALIZACAO GRAFICA ####################
                 ######################################################################
```

```
### A análise será realizada em duas partes:
### PARTE 1 - Frequência de tipo (pseudopalavras)
### PARTE 2 - Frequência de ocorrência (palavras reais)

                ############################### PARTE 1 ############################
              ############### FREQUENCIA DE TIPO: PSEUDOPALAVRAS ###############
                ######################################################################

### Criar conjunto de dados das pseudopalavras ######
dados.pseudo <- dados.brutos[dados.brutos$tipo=="pseudo",
            c("participante", "ano", "estimulo",
              "fonema", "grafema",
              "letra", "ftipoescrita","ftipoescritacat")]


### Observar dados pseudopalavras por ano escolar e fonema ###
with(dados.pseudo,table(fonema, ano))

                    ############### Analise do Fonema /s/ e seus grafemas ############
                  ############################ pseudopalavras ########################

### Criar um conjunto de dados dos grafemas 'c' e 's'

dadosp.sa<- dados.pseudo %>%
 select(participante,fonema, ano, grafema,letra,
    ftipoescrita, ftipoescritacat) %>%
 filter(fonema=="sa") %>%
 filter (letra=="s"|letra=="c") %>%
 na.omit(dadosp.sa)


### Observar dados pseudopalavras por ano escolar e grafemas ###
with(dadosp.sa,table(ano, letra))


######### GRAFICO 1 - Fonema /s/ e a escrita dos grafemas <c> e <s> nas pseudopalavras ############
## Criar um grafico dos grafemas 'c' e 's' por ano escolar

ggplot(dados.pseudo%>%
```

```
    select (participante,fonema, ano, grafema,letra,
        ftipoescrita, ftipoescritacat)%>%
    filter(fonema=="sa")%>%
    filter (letra=="s"|letra=="c")%>%
    group_by(ano, letra)%>%
    summarise(n=n())%>%
    na.omit(dados.pseudo)%>%
    mutate(porcentagem=n/sum(n)),
   aes(x = ano, y= porcentagem, fill = letra)) +
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 scale_x_discrete ("", labels= c("3A"="3°ano EF", "6A"="6°ano EF",
              "9A"="9°ano EF", "EM3"="3ª série EM")) +
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_fill_manual(labels=c("c"= "<c> -freq.", "s"= "<s> +freq."), values =c("red","blue"))+
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.5),size=7)+
 labs(x = 'Ano escolar', y = '% de escrita de grafemas', fill = 'Grafemas:') +
 theme_bw(base_size = 12) +
 theme (legend.position = "top", legend.text=
     element_text(size=25),axis.text.x = element_text(size =
                         25,color="black"),
    axis.text.y=element_blank(),
    axis.ticks=element_blank(),text = element_text (size=25, family="serif"))


### Observar número absoluto de dados em relação ao grafema por ano escolar
with(dadosp.sa,table(ano, letra))


#### Realizar Qui-quadrado da diferença entre <s> e <c> em diferentes anos escolares
escrita.grafemas.sa <- c(47, 79, 61, 85)
# Calcular Qui-quadrado
chisq.test(escrita.grafemas.sa)


######### GRAFICO 2 - Fonema /s/ e a escrita de outros grafemas nas pseudopalavras ############
## Criar um gráfico de outros grafemas por ano escolar

ggplot(dados.pseudo%>%
    select (participante,fonema, ano, grafema,letra,
```

```
        ftipoescrita, ftipoescritacat)%>%
    filter(fonema=="sa")%>%
    filter (letra != 's' & letra != 'c')%>%
    group_by(ano, letra)%>%
    summarise(n=n())%>%
    na.omit(dados.pseudo)%>%
    mutate(porcentagem=n/sum(n)),
   aes(x = ano, y= porcentagem, fill = letra)) +
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 scale_x_discrete ("", labels= c("3A"="3°ano EF", "6A"="6°ano EF",
              "9A"="9°ano EF", "EM3"="3ª série EM"))+
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_fill_manual(labels=c("f" = "<f>", "g" = "<g>", "v" = "<v>", "z" = "<z>"),
         values =c("mediumvioletred", "tan2","purple1", "olivedrab"))+
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.5),size=7)+
 labs(x = 'Ano escolar', y = '% de escrita de grafemas', fill = 'Grafemas:') +
 theme_bw(base_size = 12)+
 theme (legend.position = "top", legend.text=
     element_text(size=25),axis.text.x = element_text(size =
                          25,color="black"),
    axis.text.y=element_blank(),
    axis.ticks=element_blank(),text = element_text (size=25, family="serif"))


### Observar número absoluto de outros grafemas no conjunto de dados Fonema /sa/
with(dados.pseudo,table(ano, letra, fonema))

                        ########### Analise do Fonema /ja/ e seus grafemas ############
                    ############################ pseudopalavras ########################

### Criar um conjunto de dados dos grafemas 'g' e 'j'

dadosp.ja<- dados.pseudo %>%
 select (participante,fonema, ano, grafema,letra,
    ftipoescrita, ftipoescritacat)%>%
 filter(fonema=="ja")%>%
 filter (letra=="g"|letra=="j")%>%
 na.omit(dados.pseudo)
```

```
#### GRAFICO 3 - Fonema /ja/ e a escrita dos grafemas <g> e <j> nas pseudopalavras #####
## Criar um grafico dos grafemas 'g' e 'j' por ano escolar

ggplot(dados.pseudo%>%
    select (participante,fonema, ano, grafema,letra,
       ftipoescrita, ftipoescritacat)%>%
    filter(fonema=="ja")%>%
    filter (letra=="g"|letra=="j")%>%
    group_by(ano, letra)%>%
    summarise(n=n())%>%
    na.omit(dados.pseudo)%>%
    mutate(porcentagem=n/sum(n)),
   aes(x = ano, y= porcentagem, fill = letra)) +
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 scale_x_discrete ("", labels= c("3A"="3°ano EF", "6A"="6°ano EF",
             "9A"="9°ano EF", "EM3"="3ª série EM"))+
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_fill_manual(labels=c("g" = "<g> +freq.", "j"= "<j> -freq."), values =c("blue","red"))+
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.5),size=7)+
 labs(x = 'Ano escolar', y = '% de escrita de grafemas', fill = 'Grafemas:') +
 theme_bw(base_size = 12) +
 theme (legend.position = "top", legend.text=
     element_text(size=25),axis.text.x = element_text(size =
                         25,color="black"),
    axis.text.y=element_blank(),
    axis.ticks=element_blank(),text = element_text (size=25, family="serif"))


### Observar número absoluto de dados em relação ao grafema por ano escolar
with(dadosp.ja,table(ano, letra))


#### Realizar Qui-quadrado da diferença entre <g> e <s> em diferentes anos escolares
escrita.grafemas.ja <- c(120, 31, 128, 32)

# Calcular Qui-quadrado
chisq.test(escrita.grafemas.ja)
```

```
######### GRAFICO 4 - Fonema /ja/ e a escrita de outros grafemas nas pseudopalavras ############

## Criar um gráfico de outros grafemas por ano escolar
ggplot(dados.pseudo%>%
    select (participante,fonema, ano, grafema,letra,
       ftipoescrita, ftipoescritacat)%>%
    filter(fonema=="ja")%>%
    filter (letra != 'g' & letra != 'j')%>%
    group_by(ano, letra)%>%
    summarise(n=n())%>%
    na.omit(dados.pseudo)%>%
    mutate(porcentagem=n/sum(n)),
   aes(x = ano, y= porcentagem, fill = letra)) +
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 scale_x_discrete ("", labels= c("3A"="3°ano EF", "6A"="6°ano EF",
             "9A"="9°ano EF", "EM3"="3ª série EM"))+
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_fill_manual(labels=c("ch"="<ch>","d"="<d>", "qu"="<qu>", "x"="<x>"),
        values =c("mediumvioletred", "tan2","purple1", "olivedrab"))+
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.5),size=7)+
 labs(x = 'Ano escolar', y = '% de escrita de grafemas', fill = 'Grafemas:') +
 theme_bw(base_size = 12) +
 theme (legend.position = "top", legend.text=
     element_text(size=25),axis.text.x = element_text(size =
                        25,color="black"),
    axis.text.y=element_blank(),
    axis.ticks=element_blank(),text = element_text (size=25, family="serif"))


### Observar número absoluto de outros grafemas no conjunto de dados Fonema /sa/

with(dados.pseudo,table(ano, letra, fonema))


                ################ Analise do Fonema /za/ e seus grafemas ############
               ############################ pseudopalavras ########################
```

```
### Criar um conjunto de dados dos grafemas 's' e 'z'
dadosp.za<- dados.pseudo %>%
 select (participante,fonema, ano, grafema,letra,
    ftipoescrita, ftipoescritacat)%>%
 filter(fonema=="za")%>%
 filter (letra=="vsv"|letra=="z")%>%
 na.omit(dados.pseudo)



#### GRAFICO 5 - Fonema /z/ e a escrita dos grafemas <VsV> e <z> nas pseudopalavras
## Criar um grafico dos grafemas 's' e 'z' por ano escolar
ggplot(dados.pseudo%>%
    select (participante,fonema, ano, grafema,letra,
       ftipoescrita, ftipoescritacat)%>%
    filter(fonema=="za")%>%
    filter (letra=="vsv"|letra=="z")%>%
    group_by(ano, letra)%>%
    summarise(n=n())%>%
    na.omit(dados.pseudo)%>%
    mutate(porcentagem=n/sum(n)),
   aes(x = ano, y= porcentagem, fill = letra)) +
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 scale_x_discrete ("", labels= c("3A"="3°ano EF", "6A"="6°ano EF",
             "9A"="9°ano EF", "EM3"="3ª série EM"))+
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_fill_manual(labels=c("<vsv> +freq", "<z> -freq"), values =c("blue","red"))+
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.5),size=7)+
 labs(x = 'Ano escolar', y = '% de escrita de grafemas', fill = 'Grafemas:') +
 theme_bw(base_size = 12) +
 theme (legend.position = "top", legend.text=
     element_text(size=25),axis.text.x = element_text(size =
                         25,color="black"),
    axis.text.y=element_blank(),
    axis.ticks=element_blank(),text = element_text (size=25, family="serif"))



### Observar número absoluto de dados em relação ao grafema por ano escolar
with(dadosp.za,table(ano, letra))
```

```
#### Realizar Qui-quadrado da diferença entre <g> e <s> no 3º e 6º anos.
escrita.grafemas.za <- c(71, 81, 81, 79)
# Calcular Qui-quadrado
chisq.test(escrita.grafemas.za)

#### Realizar Qui-quadrado da diferença entre <g> e <s> no 6º e 9º anos.
escrita.grafemas.za <- c(81, 79, 69, 89)
# Calcular Qui-quadrado
chisq.test(escrita.grafemas.za)

#### Realizar Qui-quadrado da diferença entre <g> e <s> no 9º e 3ª série.
escrita.grafemas.za <- c(69, 89, 81, 115)
# Calcular Qui-quadrado
chisq.test(escrita.grafemas.za)


## Criar um grafico de outros grafemas por ano escolar

ggplot(dados.pseudo%>%
    select (participante,fonema, ano, grafema,letra,
       ftipoescrita, ftipoescritacat)%>%
    filter(fonema=="za")%>%
    filter (letra != 'vsv' & letra != 'z')%>%
    group_by(ano, letra)%>%
    summarise(n=n())%>%
    na.omit(dados.pseudo)%>%
    mutate(porcentagem=n/sum(n)),
   aes(x = ano, y= porcentagem, fill = letra)) +
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 scale_x_discrete ("", labels= c("3A"="3°ano EF", "6A"="6°ano EF",
             "9A"="9°ano EF","EM3"="3ª série EM"))+
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_fill_manual(labels=c("<j>", "<m>", "<r>", "<ss>", "<v>"),
        values =c( "brown", "tan2","purple1", "olivedrab", "mediumvioletred"))+
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.3,hjust=0.4,colour="black",size=4,
     position = position_dodge(width = 0.5)) +
 labs(x = 'Ano escolar', y = '% de ocorrência', fill = 'Grafemas:') +
```

```
 theme_bw(base_size = 12) +
 theme (legend.position = "top", legend.text=
     element_text(size=25),axis.text.x = element_text(size =
                          25,color="black"),
    axis.text.y=element_blank(),
    axis.ticks=element_blank(),text = element_text (size=25, family="serif"))


### Observar número absoluto de outros grafemas no conjunto de dados Fonema /sa/
with(dados.pseudo,table(ano, letra, fonema))


                    ###################### PARTE 1 - Pseudopalavras ####################
                        ################ Criacao e ajustes de modelos ################

####### MODELO GENERALIZADO MULTIPLO GERAL PARA AS PSEUDOPALAVRAS####
#Aqui nós vamos utilizar o conjunto de dados chamado de dados. pseudo
#repetimos o código deste conjunto abaixo para caso não esteja carregado

dados.pseudo <- dados.brutos[dados.brutos$tipo=="pseudo",
             c("participante", "ano", "estimulo",
              "fonema", "grafema",
              "letra", "ftipoescrita","ftipoescritacat")]

#Criando modelos mistos gerais para as pseudopalavras

mmpgeral.zero <-glmer(ftipoescritacat ~ 1+(1|participante), dados.pseudo,family=binomial) #modelo sem variável independente
mmpgeral.grafema <- glmer(ftipoescritacat~grafema +(1|participante), dados.pseudo, family = binomial) #modelo com grafema como v.independente.
mmpgeral.ano <- glmer(ftipoescritacat~ano+(1|participante), dados.pseudo, family = binomial) #modelo com ano como independente.
mmpgeral.grafema.ano <-glmer(ftipoescritacat~grafema+ano+(1|participante), dados.pseudo, family = binomial) #modelo com grafema + ano como v. independente.
mmpgeral.grafemaano.interacao <-glmer(ftipoescritacat~grafema*ano+(1|participante), dados.pseudo, family = binomial)#modelo com grafema em interação com ano

#ver resultados dos modelos
summary(mmpgeral.grafema) ### melhor modelo
summary(mmpgeral.ano)
```

```
summary(mmpgeral.grafema.ano)
summary(mmpgeral.grafemaano.interacao)


#fazendo comparacao de modelos aninhados
anova(mmpgeral.zero, mmpgeral.grafema) #modelo com grafema é mais explicativo do que sem nada
anova (mmpgeral.grafema, mmpgeral.ano) #modelo com grafema ano não é melhor do que com grafema.
anova (mmpgeral.grafema, mmpgeral.grafema.ano) #modelo com grafema +ano não é melhor do que com grafema)
anova (mmpgeral.grafema,mmpgeral.grafemaano.interacao) #modelo com grafema em interação com ano é melhor do que só com grafema

# O melhor modelo ajustado parece ser o seguinte
mmpgeral.grafemaano.interacao <-glmer(ftipoescritacat~grafema*ano+(1|participante), dados.pseudo, family = binomial)#modelo com grafema em interação com ano

#trabalharemos com ele daqui em diante

#grafico do modelo
drop1(mmpgeral.grafemaano.interacao ,test="Chisq")

plot(allEffects(mmpgeral.grafemaano.interacao),type = "response") #o grafico mostra a probabilidade de ocorrer o padrao de menor freq por ano e por grafema.
summary(mmpgeral.grafemaano.interacao)

#analise post-hoc
post.hoc.mmp.geral<-emmeans(mmpgeral.grafemaano.interacao,~grafema*ano,type="response",adjust="sidak")
pairs(post.hoc.mmp.geral) #aqui compara um por um
post.hoc.mmp.geral #aqui a gente consegue ver a probabilidade de ocorrer a menor freq por ano - essa prob eh a mesma do grafico abaixo.

####### GRAFICO 6 - Probabilidade de escrita do grafema com a menor frequência de tipo
#### Gráfico do modelo
grafico.modelo.geral <-plot(post.hoc.mmp.geral, color="black") +
 coord_flip()+
 labs(y= "ano", x="Probabilidade estimada de escrita do padrão com menor f.tipo")+
 facet_wrap(~grafema)+
 theme_bw()
```

```
################################ PARTE 2 ###################################
############### FREQUÊNCIA DE OCORRÊNCIA: PALAVRAS REAIS ###############
############################################################################

### Criar conjuto de dados das palavras reais ######
dados.real <- dados.brutos[dados.brutos$tipo=="real",
            c("participante", "ano", "estimulo",
              "focorrenciacat", "fonema", "grafema",
              "ftipografema", "ftipogracat","escrita", "letra",
              "certoerrado", "ftipoescrita","ftipoescritacat")]



##### GRAFICO 7 – Erros e acertos na escrita das palavras reais
## Criar grafico de erro e acerto por ano escolar

ggplot(dados.real%>%
    select (ano, fonema, grafema,letra, certoerrado)%>%
    group_by(ano, certoerrado)%>%
    summarise(n=n())%>%
    na.omit(dados.real)%>%
    mutate(porcentagem=n/sum(n)),
   aes(x = ano, y= porcentagem, fill = certoerrado)) +
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 scale_x_discrete ("", labels= c("3A"="3°ano EF", "6A"="6°ano EF",
              "9A"="9°ano EF", "EM3"="3ª série EM"))+
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_fill_manual(labels=c("certo", "errado"), values =c("green3","orange"))+
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.5),size=7)+
 labs(x = 'Ano escolar', y = '% de erros e acertos', fill = 'Tipo:') +
 theme_bw(base_size = 12) +
 theme (legend.position = "top", legend.text=
     element_text(size=25),axis.text.x = element_text(size =
                  25,color="black"),
    axis.text.y=element_blank(),
    axis.ticks=element_blank(),text = element_text (size=25, family="serif"))

#### Observar número de erro e acerto por ano escolar
with(dados.real,table(ano, certoerrado))
```

```
#### Realizar Qui-quadrado da diferença entre <g> e <s> no 9º e 3ª série.
escrita.erroseacertos <- c(409, 71, 421, 59)

# Calcular Qui-quadrado
chisq.test(escrita.erroseacertos)

                    ################ Analise do Fonema /s/ e seus grafemas ############
                   ############################ palavras reais ########################

### Criar um conjunto de dados de palavras com o Fonema /sa/
dadosr.sa <- dados.real %>%
 select (participante,ano, estimulo, focorrenciacat, fonema, grafema,
    ftipografema, ftipogracat, letra,
    ftipoescrita, ftipoescritacat, certoerrado)%>%
 filter(fonema=="sa")%>%
 na.omit(dados.real)



#### GRAFICO 8 – Fonema /s/ e os erros na escrita das palavras reais com <c> e <s>
## Criar um gráfico por ano escolar em relação ao erro
### e a frequência de ocorrência

ggplot(dadosr.sa%>%
    select (participante,ano, estimulo, focorrenciacat, fonema, grafema,
       ftipografema, letra,
       ftipoescrita, ftipoescritacat, certoerrado) %>%
      filter (focorrenciacat=="maior"|focorrenciacat=="menor")%>%
       group_by(ano, focorrenciacat,certoerrado)%>%
    summarise(n=n())%>%
    na.omit(dadosr.sa)%>%
    mutate(porcentagem=n/sum(n))%>%
    filter(certoerrado == "errado"),
   aes(x = ano, y= porcentagem, fill = focorrenciacat)) +
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 scale_x_discrete ("", labels= c("3A"="3°ano EF", "6A"="6°ano EF",
             "9A"="9°ano EF", "EM3"="3ª série EM"))+
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_fill_manual(labels=c("maior", "menor"), values =c("darkgreen","deeppink4"))+
```

```
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.5),size=7)+
 labs(x = 'Ano escolar', y = '% de erros', fill = 'Frequência de ocorrência:') +
 theme_bw(base_size = 12)+
 theme (legend.position = "top", legend.text=
     element_text(size=25),axis.text.x = element_text(size =
                         25,color="black"),
    axis.text.y=element_blank(),
    axis.ticks=element_blank(),text = element_text (size=25, family="serif"))



#### Observar número de erro no fonema /sa/
with(dadosr.sa,table(ano, focorrenciacat, certoerrado))

#### Realizar Qui-quadrado da diferença entre as palavras de maior e menor frequencia de ocorrencia.
escrita.erros.sa <- c(11, 5, 0, 0, 40, 33, 29, 24)

# Calcular Qui-quadrado
chisq.test(escrita.erros.sa)

#### GRAFICO 9 - Erros ortográficos por palavras com <c> e <s>
###Criar um grafico de erro ortografico por palavra com fonema /sa/

ggplot(dadosr.sa%>%
     select (participante,ano, estimulo, focorrenciacat, fonema, grafema,
        ftipografema, letra,
        ftipoescrita, ftipoescritacat, certoerrado) %>%
     group_by(estimulo,focorrenciacat,certoerrado)%>%
     summarise(n=n())%>%
     na.omit(dadosr.sa)%>%
     mutate(porcentagem=n/sum(n))%>%
     filter(certoerrado=="errado"),
    aes(x=estimulo,y=porcentagem, fill=focorrenciacat))+
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_fill_manual(values =c("darkgreen","deeppink4"))+
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.4),size=7)+
 labs(x = 'Palavra', y = '% de erros', fill = 'Frequência de ocorrência:') +
 theme_bw(base_size = 12)+
```

```
 theme (legend.position = "top", legend.text=
      element_text(size=25),axis.text.x = element_text(size =
                          25,color="black"),
     axis.text.y=element_blank(),
     axis.ticks=element_blank(),text = element_text (size=25, family="serif"))


#### Observar número de erro no fonema /sa/
with(dadosr.sa,table(estimulo, certoerrado))


#### Realizar Qui-quadrado da diferença entre as palavras de maior e menor frequencia de ocorrencia.
escrita.erros.sa <- c(44, 40, 18, 24, 8, 1, 7, 0)

# Calcular Qui-quadrado
chisq.test(escrita.erros.sa )

#### Realizar Qui-quadrado da diferença entre as palavras de menor frequência de ocorrência
escrita.erros.menor.sa <- c(40, 18, 22, 24)

# Calcular Qui-quadrado
chisq.test(escrita.erros.menor.sa)


## Criar um gráfico com frequência de tipo e de ocorrência por ano e por palavra
 palavrasa.coms <-ggplot(dadosr.sa%>%
                 select (participante,ano, estimulo, focorrenciacat, fonema, grafema,
                    ftipografema, letra,
                    ftipoescrita, ftipoescritacat, certoerrado) %>%
                 filter (grafema=="s")%>%
                 group_by(ano,focorrenciacat,estimulo,certoerrado)%>%
                 summarise(n=n())%>%
                 na.omit(dadosr.sa)%>%
                 mutate(porcentagem=n/sum(n))%>%
                 filter(certoerrado == "errado"),
                aes(x=ano,y=porcentagem,fill=estimulo))+
                geom_col(position="dodge",width = 0.3, colour="black")+
                 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
                 scale_x_discrete ("", labels= c("3A"="3ºano EF", "6A"="6º ano EF",
                  "9A"="9º ano EF", "EM3"="3ª série EM"))+
```

```
 scale_fill_manual(values =c("peachpuff","deeppink4", "green2"))+
 facet_wrap(~focorrenciacat)+
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.5),size=6)+
 labs(x = 'Ano escolar', y = '% de erros', fill = '') +
 theme_bw(base_size = 12)+
 theme (legend.position = "top", legend.text=
    element_text(size=25),axis.text.x = element_text(size =
                          24,color="black"),
   axis.text.y=element_blank(),
   axis.ticks=element_blank(),text = element_text (size=24, family="serif"))


palavrasa.comc <-ggplot(dadosr.sa%>%
           select (participante,ano, estimulo, focorrenciacat, fonema, grafema,
              ftipografema, letra,
              ftipoescrita, ftipoescritacat, certoerrado) %>%
           filter (grafema=="c")%>%
           group_by(ano,focorrenciacat,estimulo,certoerrado)%>%
           summarise(n=n())%>%
           na.omit(dadosr.sa)%>%
           mutate(porcentagem=n/sum(n))%>%
           filter(certoerrado == "errado"),
          aes(x=ano,y=porcentagem,fill=estimulo))+
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 scale_x_discrete ("", labels= c("3A"="3º ano EF", "6A"="6º ano EF",
              "9A"="9º ano EF", "EM3"="3ª série EM"))+
 scale_fill_manual(values =c("peachpuff","darkgreen", "deeppink4", "green2")) +
  ##scale_fill_manual(values =c("darkgreen","peachpuff","red3","deeppink4"))+
 facet_wrap(~focorrenciacat)+
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.5),size=6)+
 labs(x = 'Ano escolar', y = '% de erros', fill = '') +
 theme_bw(base_size = 12)+
 theme (legend.position = "top", legend.text=
    element_text(size=24),axis.text.x = element_text(size =
                          24,color="black"),
   axis.text.y=element_blank(),
   axis.ticks=element_blank(),text = element_text (size=24, family="serif"))
```

```
##### GRAFICO 10 - Frequência de tipo de <c> e <s> e frequência de ocorrência por ano escolar e por palavras com erros ortográficos
palavrasa.comc +palavrasa.coms


                ############### Analise do Fonema /ja/ e seus grafemas ############
                ########################### palavras reais ########################

### Criar um conjunto de dados de palavras com o Fonema /ja/
dadosr.ja <- dados.real %>%
 select (participante,ano, estimulo, focorrenciacat, fonema, grafema,
    ftipografema,ftipogracat, letra,
    ftipoescrita, ftipoescritacat, certoerrado)%>%
 filter(fonema=="ja")%>%
 na.omit(dados.real)


### GRAFICO 11 – Fonema /ʒ/ e os erros na escrita das palavras reais com <g> e <j>
## Criar um grafico por ano escolar em relação ao erro ortografico
### e a frequencia de ocorrencia

ggplot(dadosr.ja%>%
    select (participante,ano, estimulo, focorrenciacat, fonema, grafema,
       ftipografema, letra,
       ftipoescrita, ftipoescritacat, certoerrado) %>%
    filter (focorrenciacat=="maior"|focorrenciacat=="menor")%>%
    group_by(ano, focorrenciacat,certoerrado)%>%
    summarise(n=n())%>%
    na.omit(dadosr.ja)%>%
    mutate(porcentagem=n/sum(n))%>%
   filter(certoerrado == "errado"),
   aes(x = ano, y= porcentagem, fill = focorrenciacat)) +
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 scale_x_discrete ("", labels= c("3A"="3ºano EF", "6A"="6ºano EF",
             "9A"="9ºano EF", "EM3"="3ª série EM"))+
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_fill_manual(labels=c("maior", "menor"), values =c("darkgreen","deeppink4"))+
```

```
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.5),size=7)+
 labs(x = 'Ano escolar', y = '% de erros', fill = 'Frequência de ocorrência:') +
 theme_bw(base_size = 12) +
 theme (legend.position = "top", legend.text=
     element_text(size=25),axis.text.x = element_text(size = 25,color="black"),
    axis.text.y=element_blank(),
    axis.ticks=element_blank(),text = element_text (size=25, family="serif"))


#### Observar número de erro no fonema /ja/
with(dadosr.ja,table(ano, focorrenciacat, certoerrado))


#### Realizar Qui-quadrado da diferença entre as palavras de maior e menor frequencia de ocorrencia.
escrita.erros.ja <- c(25, 17, 6, 8, 30, 22, 15, 10)

# Calcular Qui-quadrado
chisq.test(escrita.erros.ja)


### GRAFICO 12 – Erros ortográficos por palavras com <g> e <j>
###Criar um gráfico de erro ortográfico por palavra com fonema /ja/

ggplot(dadosr.ja%>%
    select (participante,ano, estimulo, focorrenciacat, fonema, grafema,
       ftipografema, letra,
       ftipoescrita, ftipoescritacat, certoerrado) %>%
    group_by(estimulo,focorrenciacat,certoerrado)%>%
    summarise(n=n())%>%
    na.omit(dadosr.ja)%>%
    mutate(porcentagem=n/sum(n))%>%
    filter(certoerrado=="errado"),
   aes(x=estimulo,y=porcentagem, fill=focorrenciacat))+
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_fill_manual(values =c("darkgreen","deeppink4"))+
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.5),size=7)+
 labs(x = 'Palavras', y = '% de erros', fill = 'Frequência de ocorrência:') +
```

```
 theme_bw(base_size = 12)+
 theme (legend.position = "top", legend.text=
     element_text(size=25),axis.text.x = element_text(size =
                         25,color="black"),
    axis.text.y=element_blank(),
    axis.ticks=element_blank(),text = element_text (size=25, family="serif"))



#### Observar número de erro no fonema /ja/ por palavra
with(dadosr.ja,table(estimulo, certoerrado))

#### Realizar Qui-quadrado da diferença entre as palavras de maior e menor frequencia de ocorrencia.
escrita.erros.ja.palavras <- c(45, 2, 6, 31, 2, 33, 21, 0)

# Calcular Qui-quadrado
chisq.test(escrita.erros.ja.palavras)

## Criar um grafico com frequencia de tipo e de ocorrencia por ano e por palavra

palavrasa.comg <-ggplot(dadosr.ja%>%
           select (participante,ano, estimulo, focorrenciacat, fonema, grafema,
              ftipografema, letra,
              ftipoescrita, ftipoescritacat, certoerrado) %>%
           filter (grafema=="g")%>%
           group_by(ano,focorrenciacat,estimulo,certoerrado)%>%
           summarise(n=n())%>%
           na.omit(dadosr.ja)%>%
           mutate(porcentagem=n/sum(n))%>%
           filter(certoerrado == "errado"),
          aes(x=ano,y=porcentagem,fill=estimulo))+
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 scale_x_discrete ("", labels= c("3A"="3º ano EF", "6A"="6º ano EF",
              "9A"="9º ano EF", "EM3"="3ª série EM"))+
 scale_fill_manual(labels=c("general", "gestual","gincana"),values =c("darkgreen","peachpuff","deeppink4"))+
 facet_wrap(~focorrenciacat)+
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.5),size=7)+
 labs(x = 'Ano escolar', y = '% de erros', fill = '') +
 theme_bw(base_size = 12)+
```

```
 theme (legend.position = "top", legend.text=
     element_text(size=25),axis.text.x = element_text(size =
                         25,color="black"),
    axis.text.y=element_blank(),
    axis.ticks=element_blank(),text = element_text (size=25, family="serif"))

palavrasa.comj <-ggplot(dadosr.ja%>%
            select (participante,ano, estimulo, focorrenciacat, fonema, grafema,
               ftipografema, letra,
               ftipoescrita, ftipoescritacat, certoerrado) %>%
            filter (grafema=="j")%>%
            group_by(ano,focorrenciacat,estimulo,certoerrado)%>%
            summarise(n=n())%>%
            na.omit(dadosr.ja)%>%
            mutate(porcentagem=n/sum(n))%>%
            filter(certoerrado == "errado"),
           aes(x=ano,y=porcentagem,fill=estimulo))+
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 scale_x_discrete ("", labels= c("3A"="3º ano EF", "6A"="6º ano EF",
              "9A"="9º ano EF", "EM3"="3ª série EM"))+
 scale_fill_manual(labels= c("dejeto", "jiló", "jipe", "sujeito"), values =c( "peachpuff",
"deeppink4","darkgreen","green2"))+
 facet_wrap(~focorrenciacat)+
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.5),size=7)+
 labs(x = 'Ano escolar', y = '% de erros', fill = '') +
 theme_bw(base_size = 12)+
 theme (legend.position = "top", legend.text=
     element_text(size=25),axis.text.x = element_text(size =
                         25,color="black"),
    axis.text.y=element_blank(),
    axis.ticks=element_blank(),text = element_text (size=25, family="serif"))


#### GRAFICO 13 - Frequência de tipo de <g> e <j>
### e frequência de ocorrência por ano escolar e por palavras com erros ortográficos
palavrasa.comg +palavrasa.comj
```

```
                ############### Analise do Fonema /za/ e seus grafemas ############
                ########################### palavras reais ########################

### Criar um conjunto de dados de palavras com o Fonema /za/
dadosr.za <- dados.real %>%
 select (participante,ano, estimulo, focorrenciacat, fonema, grafema,
    ftipografema, letra,
    ftipoescrita, ftipoescritacat, certoerrado)%>%
 filter(fonema=="za")%>%
 na.omit(dados.real)


### GRAFICO 14 – Fonema /z/ e os erros na escrita das palavras reais com <s> e <z>
## Criar um grafico por ano escolar em relação ao erro ortografico
### e a frequencia de ocorrencia

ggplot(dadosr.za%>%
    select (participante,ano, estimulo, focorrenciacat, fonema, grafema,
       ftipografema, letra,
       ftipoescrita, ftipoescritacat, certoerrado) %>%
    filter (focorrenciacat=="maior"|focorrenciacat=="menor")%>%
    group_by(ano, focorrenciacat,certoerrado)%>%
    summarise(n=n())%>%
    na.omit(dadosr.za)%>%
    mutate(porcentagem=n/sum(n)) %>%
   filter(certoerrado =="errado"),
   aes(x = ano, y= porcentagem, fill = focorrenciacat)) +
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 scale_x_discrete ("", labels= c("3A"="3°ano EF", "6A"="6°ano EF",
             "9A"="9°ano EF", "EM3"="3ª série EM"))+
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_fill_manual(labels=c("maior", "menor"), values =c("darkgreen","deeppink4"))+
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.5),size=7)+
 labs(x = 'Ano escolar', y = '% de erros', fill = 'Frequência de ocorrência:') +
 theme_bw(base_size = 12) +
 theme (legend.position = "top", legend.text=
     element_text(size=25),axis.text.x = element_text(size =
                    25,color="black"),
```

```
    axis.text.y=element_blank(),
    axis.ticks=element_blank(),text = element_text (size=25, family="serif"))


#### Observar número de erro no fonema /za/
with(dadosr.za,table(ano, focorrenciacat, certoerrado))


#### Realizar Qui-quadrado da diferença entre as palavras de maior e menor frequencia de ocorrencia.
escrita.erros.Za <- c(10, 12, 1, 2, 34, 24, 20, 14)

# Calcular Qui-quadrado
chisq.test(escrita.erros.Za)


##### GRAFICO 15 – Erros ortográficos por palavras com <s> e <z>
#Criar um grafico de erro ortografico por palavra com fonema /za/

ggplot(dadosr.za%>%
    select (participante,ano, estimulo, focorrenciacat, fonema, grafema,
       ftipografema, letra,
       ftipoescrita, ftipoescritacat, certoerrado) %>%
    group_by(estimulo,focorrenciacat,certoerrado)%>%
    summarise(n=n())%>%
    na.omit(dadosr.za)%>%
    mutate(porcentagem=n/sum(n))%>%
    filter(certoerrado=="errado"),
   aes(x=estimulo,y=porcentagem, fill=focorrenciacat))+
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_fill_manual(values =c( "darkgreen","deeppink4"))+
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.5),size=7)+
 labs(x = 'palavra', y = '% de erros', fill = 'Frequência de ocorrência:') +
 theme_bw(base_size = 12)+
 theme (legend.position = "top", legend.text=
     element_text(size=25),axis.text.x = element_text(size =
                        25,color="black"),
    axis.text.y=element_blank(),
    axis.ticks=element_blank(),text = element_text (size=25, family="serif"))
```

```
#### Observar número de erro no fonema /za/ por palavra
with(dadosr.za,table(estimulo, certoerrado))


#### Realizar Qui-quadrado da diferença entre as palavras de maior e menor frequencia de ocorrencia.
escrita.erros.za.palavras <- c(18, 22, 22, 30, 5, 9, 2, 9)

# Calcular Qui-quadrado
chisq.test(escrita.erros.za.palavras)


#### Gráfico 16 - Frequência de tipo de <s> e <z> e frequência de ocorrência por ano escolar
### e por palavras com erros ortográficos
## Criar um grafico com frequencia de tipo e de ocorrencia por ano e por palavra

palavraza.coms <-ggplot(dadosr.za%>%
          select (participante,ano, estimulo, focorrenciacat, fonema, grafema,
             ftipografema, letra,
             ftipoescrita, ftipoescritacat, certoerrado) %>%
          filter (grafema=="s")%>%
          group_by(ano,focorrenciacat,estimulo,certoerrado)%>%
          summarise(n=n())%>%
          na.omit(dadosr.za)%>%
          mutate(porcentagem=n/sum(n))%>%
          filter(certoerrado == "errado"),
         aes(x=ano,y=porcentagem,fill=estimulo))+
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 scale_x_discrete ("", labels= c("3A"="3?ano EF", "6A"="6?ano EF",
             "9A"="9?ano EF", "EM3"="3? s?rie EM"))+
 scale_fill_manual(labels= c("camisa", "divisa", "surpresa", "turquesa"), values =c("darkgreen",
"deeppink4","green2","peachpuff"))+
 facet_wrap(~focorrenciacat)+
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
    vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.5),size=7)+
 labs(x = 'Ano escolar', y = '% de erros', fill = '') +
 theme_bw(base_size = 12)+
 theme (legend.position = "top", legend.text=
    element_text(size=25),axis.text.x = element_text(size =
```

```
                    25,color="black"),
    axis.text.y=element_blank(),
    axis.ticks=element_blank(),text = element_text (size=25, family="serif"))

palavrza.comz <-ggplot(dadosr.za%>%
          select (participante,ano, estimulo, focorrenciacat, fonema, grafema,
             ftipografema, letra,
             ftipoescrita, ftipoescritacat, certoerrado) %>%
          filter (grafema=="z")%>%
          group_by(ano,focorrenciacat,estimulo,certoerrado)%>%
          summarise(n=n())%>%
          na.omit(dadosr.sa)%>%
          mutate(porcentagem=n/sum(n))%>%
          filter(certoerrado == "errado"),
         aes(x=ano,y=porcentagem,fill=estimulo))+
 geom_col(position="dodge",width = 0.3, colour="black")+
 scale_y_continuous (labels=scales::percent_format(accuracy=1))+
 scale_x_discrete ("", labels= c("3A"="3º ano EF", "6A"="6º ano EF",
             "9A"="9º ano EF", "EM3"="3ª série EM"))+
 scale_fill_manual(labels=c("amizade",      "beleza","leveza","ojeriza"),values      =c("darkgreen",      "green2",
"peachpuff","deeppink4"))+
 facet_wrap(~focorrenciacat)+
 geom_text(aes(label=scales::percent(porcentagem,accuracy=1)),
     vjust=-0.2,hjust=0.4,colour="black",position=position_dodge(.5),size=7)+
 labs(x = 'Ano escolar', y = '% de erros', fill = '') +
 theme_bw(base_size = 12)+
 theme (legend.position = "top", legend.text=
     element_text(size=25),axis.text.x = element_text(size =
                    25,color="black"),
    axis.text.y=element_blank(),
    axis.ticks=element_blank(),text = element_text (size=25, family="serif"))

#### Gráfico 16 - Frequência de tipo de <s> e <z> e frequência de ocorrência por ano escolar
### e por palavras com erros ortográficos
palavraza.coms + palavrza.comz


            ################################ PALAVRAS REAIS ####################
                ###################### Criação e ajustes de modelos ################
```

```
### Criar Modelo linear generalizado misto geral para palavras reais
view(dados.real)

m.real.zero <- glmer(certoerrado ~ 1+(1|participante), dados.real,family=binomial)
m.real.ano <- glmer(certoerrado ~ ano+(1|participante), dados.real,family=binomial)
m.real.ocorrencia <- glmer(certoerrado ~ focorrenciacat+(1|participante), dados.real,family=binomial)
m.real.tipo <- glmer(certoerrado ~ ftipogracat+(1|participante), dados.real,family=binomial)
m.real.ano.ocorrencia <- glmer(certoerrado ~ ano+focorrenciacat + (1|participante), dados.real,family=binomial)
m.real.ano.ocorrencia.tipo <- glmer(certoerrado ~ ano+focorrenciacat+ftipogracat+(1|participante), dados.real,family=binomial)
m.interacao1 <- glmer(certoerrado ~ ano + focorrenciacat*ftipogracat+(1|participante), dados.real,family=binomial)
m.interacao2 <- glmer(certoerrado ~ ano*focorrenciacat*ftipogracat+(1|participante), dados.real,family=binomial)

#ver resultados dos modelos
summary(m.real.ano)
summary (m.real.ocorrencia)
summary (m.real.tipo)
summary(m.real.ano.ocorrencia)
summary(m.real.ano.ocorrencia.tipo)
summary (m.interacao)


#fazendo comparacao de modelos aninhados
anova(m.real.zero, m.real.ano) #modelo com ano é melhor do que sem nada.
anova (m.real.ano, m.real.ocorrencia) # modelo com ano não difere do modelo com ocorrência
anova (m.real.ano, m.real.tipo) #modelo com ano não difere do modelo com tipo
anova (m.real.ano, m.real.ano.ocorrencia) #modelo com ano e ocorrência é melhor do que só com ano
anova (m.real.ano.ocorrencia, m.real.ano.ocorrencia.tipo) #modelo com ano, ocorrência e tipo é melhor do que só com ocorrencia e ano.
anova(m.real.ano.ocorrencia.tipo,m.interacao1) #interacao entre ocorrência e tipo não é melhor.
anova (m.real.ano.ocorrencia.tipo, m.interacao2) #nao tem interação

## o melhor modelo é o seguinte
m.real.ano.ocorrencia.tipo <- glmer(certoerrado ~ ano+focorrenciacat+ftipogracat+(1|participante), dados.real,family=binomial)

# este modelo indica que os indices de erros ortograficos são explicados pelo ano, pela frequencia de ocorrencia e pela frequencia de tipo
```

```
#trabalharemos com ele daqui em diante

#grÁfico do modelo
drop1(m.real.ano.ocorrencia.tipo,test="Chisq")

plot(allEffects(m.real.ano.ocorrencia.tipo),type = "response") #o grafico mostra a probabilidade de ocorrer o padrao de menor freq por ano e por grafema.

summary(m.real.ano.ocorrencia.tipo)

#analise post-hoc
post.hoc.m.real<-emmeans(
m.real.ano.ocorrencia.tipo,~ano+focorrenciacat+ftipogracat,type="response",adjust="sidak")
pairs(post.hoc.m.real) #aqui compara um por um
post.hoc.m.real #aqui conseguimos ver a probabilidade de ocorrer a menor freq por ano - essa prob eh a mesma do grafico abaixo.


##### GRAFICO 17 - Probabilidade de erro ortográfico em palavras reais
grafico.modelo.geral <-plot(post.hoc.m.real, color="black") +
 coord_flip()+
 labs(x="probabilidade estimada de erros ortográficos em palavras reais")+
 #facet_wrap(~focorrenciacat)+
 theme_bw(base_size = 12)
```