DOI: https://doi.org/10.18309/ranpoll.v56.2016
v. 56, e2016, 2025

revista da anpoll
ISSN 1982-7830

DOSSIÊ

Inovação, Tecnologia e Inclusão: Corpora aplicados ao estudo da dislexia e a proposta do DYSCORP (Dyslexia *Corpus*)

Innovation, Technology and Inclusion: Corpora applied to the study of dyslexia and the DYSCORP (Dyslexia *Corpus*) proposal

Aline Pacheco
Aline Evers
Aline Fay de Azevedo

Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, RS, Brasil.
E-mails: aline.pacheco@pucrs.br; aline.evers@pucrs.br; aline.azevedo@pucrs.br

RESUMO: Pesquisas recentes nas áreas de Neurociência, Psicologia Cognitiva e Linguística têm aprimorado a compreensão dos processos neurais da leitura. No entanto, muitos desses achados não chegam efetivamente às salas de aula e aos consultórios, mesmo com o aumento de diagnósticos de transtornos do neurodesenvolvimento, como a dislexia, que afetam leitura e escrita. *Corpora* específicos para estudos sobre a dislexia não são abundantes e trazem informações de sujeitos falantes de línguas como inglês, árabe e espanhol, e não com falantes de português brasileiro. Este artigo apresenta a proposta inicial do DYSCORP (*Dyslexia Corpus*), um banco de dados para investigar a dislexia sob a ótica da Linguística de *Corpus*. Constituído de subcorpora, com listas de palavras e erros típicos da dislexia, o DYSCORP permite análises linguísticas diversas. Os dados cobrem diferentes faixas etárias: crianças e adolescentes (Projeto ACERTA) e adultos (estudantes universitários). O objetivo é construir um banco de dados robusto sobre sujeitos disléxicos falantes de português brasileiro, auxiliando na investigação, diagnóstico e intervenção precoce da dislexia. Inserido no escopo de tecnologias educacionais, o DYSCORP permitirá criar não somente material referencial na área, mas a criação de um protótipo de APP que apoie o trabalho de profissionais das áreas do ensino e da saúde.

PALAVRAS-CHAVE: Linguística de *Corpus;* Dislexia*; Corpus* para Propósitos Específicos; EdTech.

ABSTRACT: Recent research in the areas of Neuroscience, Cognitive Psychology, and Linguistics has improved our understanding of the neural processes of reading. However, it is noteworthy that many of these findings do not effectively reach classrooms and clinics, even with the increase in diagnoses of neurodevelopmental disorders, such as dyslexia, which affect reading and writing. Specific corpora for studies on dyslexia are not abundant and bring information from subjects

COMO CITAR

PACHECO, Aline; EVERS, Aline; AZEVEDO, Aline Fay de. Inovação, Tecnologia e Inclusão: Corpora aplicados ao estudo da dislexia e a proposta do DYSCORP (Dyslexia *Corpus*). *Revista da Anpoll,* v. 56, e2016, 2025. doi: https://doi.org/10.18309/ranpoll.v56.2016

RECEBIDO: 27/01/2025; ACEITO: 25/05/2025

who speak languages such as English, Arabic, and Spanish, and not from speakers of Brazilian Portuguese. This article presents the initial proposal of DYSCORP (Dyslexia *Corpus*), a database for investigating dyslexia from the perspective of *Corpus* Linguistics. It consists of subcorpora, with lists of words and errors typical of dyslexia allowing for diverse linguistic analyses. The data covers different age groups: children and adolescents (ACERTA Project) and adults (university students). The goal is to build a robust database on dyslexic subjects who speak Brazilian Portuguese, assisting in the investigation, diagnosis and early intervention of dyslexia. Inserted in the scope of educational technologies, DYSCORP will allow the creation of not only reference material in the area, but also the creation of an APP that can support the work of professionals in the areas of education and health.

KEYWORDS: *Corpus* Linguistics; Dyslexia; *Corpus* for Specific Purposes; EdTech.

1 INTRODUÇÃO

Nos últimos anos, as Humanidades Digitais têm ganhado destaque no Brasil, fomentando o uso intensivo de recursos e ferramentas digitais (Freitas, 2017). Com princípios como dados abertos, compartilhamento e interdisciplinaridade, essa abordagem é essencial para as pesquisas Linguísticas e que envolvem linguagens no século XXI. Aliada a esses princípios, no âmbito dos Estudos da Linguagem, há a exploração automática de grandes acervos textuais, em que se destaca a Linguística de *Corpus*, cujas pesquisas oferecem há décadas a constituição de acervos textuais e recursos para propósitos específicos, marcando a área da Linguística como potência de inovação, de produção de novas tecnologias e de promoção da inclusão em contextos de pesquisa e educacionais diversos.

O desenvolvimento de estratégias educacionais personalizadas, trazendo inovação para as tarefas de adaptação de materiais pedagógicos, é um movimento essencial para contornar, por exemplo, o baixo desempenho em leitura. Conforme Rodriguez-Segura (2022), o surgimento de tecnologias educacionais (EdTechs) nos países em desenvolvimento tem sido recebido como um caminho promissor para abordar algumas das questões políticas mais desafiadoras dentro dos sistemas educacionais. O autor define EdTech como qualquer aplicação de tecnologias de informação e comunicação na educação, tal como distribuição de tecnologia existente, fornecimento de dispositivos com software personalizado, adaptação de tecnologias e o uso de software especializado em computadores comuns.

Sabe-se que o baixo desempenho em leitura não é atribuído a uma única causa. Pesquisas recentes nas áreas da Neurociência, Psicologia Cognitiva e Linguística têm aprimorado a compreensão dos processos neurais da leitura (Azevedo, 2016). No entanto, é notável que resultados dessas pesquisas não chegam efetivamente às salas de aula e aos consultórios e clínicas. Ao longo dos anos, cursos de formação de professores têm perpetuado conceitos já desbancados pela ciência da leitura, como o de que "não se deve apressar a criança, cada uma tem seu tempo para aprender a ler". Sally Shaywitz (2020) critica essa abordagem, descrevendo-a como uma filosofia de "esperar para falhar" ("wait to fail"), argumentando que, do ponto de vista do desenvolvimento neural e das possibilidades de intervenção precoce, esse tipo de espera compromete as chances de recuperação da criança.

A conexão de conhecimentos de áreas correlatas é necessária para avanços na compreensão da produção de recursos disponíveis, tais como ferramentas, aplicativos, e corpora que possam ser disponibilizados, em especial, para a dislexia, tema sobre o qual nos debruçamos

neste artigo. Importa frisar que, apesar das abordagens estatísticas abundantes advindas da Linguística de *Corpus*, há participação ainda tímida de linguistas, professores e pesquisadores da área da Linguística na planificação de materiais desse cunho, que via de regra são produzidos em massa por pedagogos, psicopedagogos, cientistas da computação e engenheiros de software.

A Linguística de *Corpus*, enquanto método empírico de investigação linguística, muito tem contribuído para investigação e análise de ocorrências reais de linguagem à medida que se dedica, basicamente, à exploração de dados através de softwares especializados (Berber Sardinha, 2004; Sinclair, 1991; Pacheco *et al.*, 2023). Corpora especializados no estudo da dislexia são escassos (Pedler, 2007; Rello; Baeza-Yates; Llisterri, 2014, Alamri; Thealan, 2017) e, com sujeitos falantes de português brasileiro, não encontrados. Nesse sentido, Pacheco (2024) propõe a compilação do DYSCORP, o *Dyslexia Corpus*, com dados previamente coletados e investigados por Azevedo (2016) e à luz de taxonomias de investigação e tecnologias sugeridas em Evers (2013, 2018, 2024).

Este artigo objetiva apresentar argumentos que reforcem a necessidade de dispormos de um banco de dados robusto e organizado advindo de sujeitos disléxicos falantes de português brasileiro a fim de permitir a criação de tecnologias que atendam às demandas de otimização de efetiva inclusão de tal público no âmbito de ensino e saúde. Para tanto, o texto se apresenta da seguinte maneira: a seção dois revisa considerações básicas para o entendimento da dislexia, seus desdobramentos e impactos. A terceira seção descreve brevemente *corpora* encontrados sobre o tema. A seção 4 apresenta recursos e ferramentas já disponíveis para tratamento da dislexia e para a adaptação de materiais a sujeitos com dislexia, citando alguns exemplos a fim de reforçar o caráter inovador da proposta do DYSCORP, apresentada na seção 5. Como veremos, embora se encontre em estágio embrionário, a proposta é inovadora uma vez que traz dados organizados por linguistas e à luz da Linguística de *Corpus* e promove maior inclusão de sujeitos disléxicos através da investigação de suas principais dificuldades por linguistas, dificuldades estas igualmente analisadas por profissionais da saúde e contempladas na criação de tecnologias que facilitem acesso a informações e tratamento adequado.

2 CONSIDERAÇÕES SOBRE A DISLEXIA

O conhecimento adquirido pela leitura é relevante para a qualificação profissional e é essencial na formação de cidadãos responsáveis e engajados (Morais, 2013). Contudo, o contexto educacional brasileiro é marcado por desafios significativos no processo de alfabetização. Em 2014, o Brasil apresentava uma taxa de analfabetismo de 8,3% entre pessoas com 15 anos ou mais (IBGE, 2012). No âmbito da avaliação de competência em leitura do PISA, em 2015, o Brasil obteve a 59ª posição entre 65 países. Além disso, o número de crianças entre 6 e 7 anos incapazes de ler e escrever aumentou 66,3% no Brasil entre 2020 e 2021, em decorrência da pandemia.

Dentre os desafios encontrados no contexto educacional, para além dos dados apontados, é importante ressaltar o aumento do número de sujeitos diagnosticados com transtornos do neurodesenvolvimento, como a dislexia do desenvolvimento, que afeta a leitura e a escrita. O termo dislexia, originado do grego (*dys* = dificuldade + *lexia* = palavras), refere-se a um

transtorno que pode afetar um em cada dez indivíduos e tem se tornado relevante em discussões nas áreas de Educação, Psicologia, Neuropsicologia e Linguística (Alamri; Teahan, 2017; Azevedo, 2016; Rello, 2014; Azevedo *et al.*, 2023).

Segundo a Associação Nacional de Dislexia – AND ([20--?]), mais de 7,8 milhões de brasileiros têm dislexia, representando 4% da população. O diagnóstico e tratamento adequados oferecidos a pessoas com dislexia trazem comprovadamente melhorias significativas no desempenho acadêmico, na autoestima e no bem-estar psicológico desses indivíduos, que enfrentam desafios diversos devido à dificuldade de leitura e de escrita.

A dislexia é um transtorno específico de aprendizagem, com origem neurobiológica, caracterizado pela dificuldade no reconhecimento preciso e fluente das palavras, na decodificação e na soletração. Por ser persistente, sua identificação precoce é fundamental para evitar o agravamento das dificuldades, permitindo a intervenção adequada e a redução do risco de evasão escolar. Dificuldades no reconhecimento de palavras corretas, ortografia e habilidade de decodificação precárias são algumas das características da dislexia, resultado de um déficit no componente fonológico da linguagem (Rello; Baeza-Yates; Llisterri, 2014). Como consequência, problemas na compreensão leitora e na experiência reduzida em leitura podem prejudicar a expansão do repertório lexical e de conhecimentos gerais. Nessa linha, tentativas de investigação e análise de ocorrências lexicais são válidas à medida que permitem um acesso maior a diagnósticos e tratamentos para sujeitos com dislexia.

O projeto ACERTA, uma parceria entre a Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), o Instituto do Cérebro (InsCer) e outras instituições no Brasil, viabilizou a investigação de crianças em fase de alfabetização para identificar elementos que pudessem prever dificuldades de aprendizagem. Dentre os dados produzidos no escopo do projeto, parte dos protocolos utilizados durante a coleta de dados permanece inexplorado. Alguns dos recursos utilizados em coletas de dados para pesquisas do projeto contêm textos produzidos por indivíduos disléxicos e constituem fonte rica de dados para outras pesquisas, especialmente se organizados de modo computadorizado e padronizado por princípios que permitam análises acadêmicas apropriadas. Por isso, propomos a Linguística de *Corpus* como método que permeie a compilação desse banco de dados específico.

3 *CORPORA* ESPECIALIZADOS PARA APLICAÇÃO À DISLEXIA

Entendida prioritariamente como uma abordagem metodológica de pesquisa ao estudo da linguagem (Biber *et al.*, 2010), a Linguística de *Corpus* se propõe a estudar a linguagem em exemplos de uso real (McEnery; Wilson, 1996) a partir de um *corpus*, definido por Tagnin (2013, p. 29) como "uma coletânea de textos, necessariamente em formato eletrônico, compilados e organizados segundo critérios ditados pelo objetivo da pesquisa a que se destina".

Ainda, um *corpus* se caracteriza por sua autenticidade, forma legível para máquina, além de aspectos particularmente associados ao armazenamento e tratamento dos dados, como processos de etiquetagem. A etiquetagem é um processo de identificação do aspecto específico proposto pela pesquisa, podendo ser de natureza léxica, sintática, morfossintática, semântica, pragmática ou fonética. Tal processo pode ser realizado manualmente ou por

máquina, e permite ao pesquisador determinar o dado para o propósito desejado de modo a poder ser lido por máquina (Berber Sardinha, 2004; Pacheco, 2010). Dentre os *softwares* que tornam possível a análise de *corpora*, destacam-se alguns comumente utilizados: *WordSmith Tools*[1], AntConc[2] e *SketchEngine*[3].

Granger (2002) descreve uma ferramenta de etiquetagem de erros chamada *Error Editor*, que viabiliza a pesquisadores etiquetarem erros a partir dos textos de aprendizes a fim de, posteriormente, elaborarem listas de erros típicos e facilitarem o desenvolvimento de programas computacionais que possam ajudar no processo de aprendizagem, no que concerne à gramática, ortografia, verificação de estilo e outros. O *Error Editor* caracteriza-se por uma série de etapas, dentre elas a correção manual dos erros do aprendiz e classificação desse erro dentro de um "*error tag*" (ou etiqueta de erro). Tal etiqueta consiste num código maior e uma série de códigos menores. Após esse processo, os erros podem ser pesquisados a partir dos códigos (Granger, 2002; Pacheco, 2010).

Tal como entendido pela Linguística de *Corpus*, *corpora* compilados a partir de erros cometidos por disléxicos podem ser encontrados em diferentes línguas: em inglês (Pedler, 2007; Pedler; Milton, 2010), em espanhol (Rello, 2014; Rello; Baeza-Yates; Llisterri, 2014), em dinamarquês (Pedersen; Larsen, 2010), e em árabe (Alamri; Teahan, 2017), mas não em português brasileiro. Pedler (2007) apresenta um programa de correção ortográfica de erros em palavras utilizadas por disléxicos extraídos de textos variados, contendo aproximadamente 21.000 palavras com 2.800 erros em inglês. Num primeiro momento, a autora criou uma lista a partir de conjuntos de palavras mais provavelmente confundidas. A Tabela 1 apresenta a composição total desse *corpus*.

Tabela 1 – Composição do *corpus*

Source	Sentences	Words	Non-word errors	Word-boundary errors	Real-word errors	Total errors
Homework*	67	974	205	28	44	278
Compositions*	61	845	103	8	50	161
Office documents*	67	720	161	6	26	193
IT NVQ	48	732	20	18	51	89
Undergraduate essays	341	6382	421	8	128	557
Typing experiment	61	694	86	7	39	132
Child's stories	11	126	7	0	14	21
Bulletin boards	739	11051	678	77	468	1223
Totals	1395	21524	1681	152	820	2654

Fonte: Pedler (2007)

[1] https://www.lexically.net/wordsmith/

[2] https://www.laurenceanthony.net/software/antconc/

[3] https://www.sketchengine.eu/

A etiquetagem foi feita manualmente, mostrando a palavra correta e identificando a errada produzida pelos sujeitos, tal como na Figura 1:

The only **<ERR targ=weaknesses> weakness </ERR>** that I think I had were the **<ERR targ=material> martial </ERR>** on each slide.

Figura 1 – Modelo de etiquetagem utilizado por Pedler
Fonte: Pedler (2007)

Após a etiquetagem e através da linguagem de programação *Perl*, o *corpus* permitia acesso a dados referentes a frequência de erros específicos, de palavras mais frequentes, de conjuntos de palavras mais comumente confundidos, informações que foram usadas para a criação do programa de correção automática.

Rello (2014) discorre sobre a construção de exercícios com palavras para crianças disléxicas a partir do *Dyscorpus*, um *corpus* em espanhol fundamentado em textos escritos por crianças disléxicas entre 6 e 15 anos, contendo 83 textos – 54 de trabalhos escolares e 29 oriundos dos pais. O *corpus* tem um total de 1.171 erros e é anotado a partir de uma lista de erros manualmente extraída a partir das seguintes informações:

- características dos erros envolvidos: simples múltiplos, erros ortográficos que não resultam em outra palavra correta; erros reais das palavras, ou seja, que resultam em outra palavra correta; tipo de erros, tais como substituição, inserção, omissão e transposição; erros nos finais das palavras; erros de primeira letra;
- características envolvidas nos erros;
- contexto dos erros.

Desdobrando essa linha de pesquisa, Rello; Baeza-Yates; Llisterri, (2014) salientam a dificuldade de encontrar dados linguísticos de pessoas com dislexia e apresentam o *Dyslist*, um *corpus* em espanhol composto de uma lista com erros coletada de 83 textos escritos por disléxicos. Tais erros foram anotados a partir de características fonéticas e visuais, somando 887 palavras erradas. As principais informações usadas nesse processo foram:

- palavra-alvo;
- palavra escrita de forma errada;
- a distância "*Damerau-Levenshtein*": número mínimo de correções referentes à inserção, apagamento, substituição, transposição;
- frequência da palavra-alvo e da palavra errada;
- número de caracteres da palavra-alvo e da palavra errada;
- posição do erro na palavra-alvo;
- número de sílabas e estrutura silábica da palavra-alvo;
- tipo de erro (substituição, inserção, apagamento, transposição)
- "palavra-real", ou seja, se o erro acaba produzindo uma outra palavra que já existe;
- informação visual;

- letra espelhada;
- informação fonética tipo de fone;
- transferência linguística.

Pedersen e Larsen (2010) apresentam um *corpus* especializado na língua dinamarquesa em investigação sobre o discurso de indivíduos disléxicos comparado à performance em ASR (*Automatic speech recognition*, ou reconhecimento automático de discurso) com o objetivo de permitir que tais sujeitos possam otimizar suas habilidades de leitura de modo independente (sem a supervisão de um profissional) através de um tutor automático para leitura. A justificativa é a de que materiais de suporte para dislexia, em sua maioria, concentram-se em crianças e em sistemas de aquisição de segunda língua e, para isso, compilam um *corpus* a partir de gravações de leitura performada por oito adultos disléxicos, a partir de material escrito aprovado por terapeutas, totalizando cem minutos de gravação e 7.578 palavras.

A anotação se fundamentou nas seguintes categorias:

Tabela 2 – Categorias de Anotação do *Corpus*

Categoria	Conteúdo
"Ortho"	Anotação ortográfica, no nível da palavra
"Error"	Tipo de erro (inserção, substituição, omissão, separação)
"Prompt"	Palavra "certa" a ser lida
"Timing Event"	Inconsistência temporal (pausas ou "pulos" na leitura")

Fonte: Autoras, adaptado de Pedersen e Larsen (2010)

A análise do *corpus* gerou dados para a criação de uma configuração de reconhecimento baseada em ASR que permite a detecção de problemas de leitura em discurso disléxico.

Alamri e Thealan (2017) apresentam uma nova categoria de anotação para o *corpus* BDAC *Corpus* (*Arabic dyslexia corpus*), uma classificação de erros disléxicos para textos em árabe denominada DECA (*dyslexic error classification for Arabic texts*). Essa classificação tem por base uma lista de erros ortográficos extraídos de estudos anteriores e um compilado de textos produzidos por disléxicos que fornecem um modelo de análise com 37 tipos de erros agrupados em nove categorias, a saber: Hamza, Almadd, Confusão, Diacríticos, Forma, Erros Comuns, Diferenças, Método de Escrita, Letra escrita mas não pronunciada – ou vice-versa – e outros. Os autores atentam para peculiaridades da língua árabe, ilustrada pelo nome das duas primeiras categorias.[4]

O *Corpus* BDAC tem 27.136 palavras e 8.000 erros de textos produzidos por crianças de 8 a 12 anos, de gêneros masculino e feminino, diagnosticadas com dislexia por profissionais especializados. Os textos foram coletados em escolas primárias na Arábia Saudita, em formulários *on-line* e textos fornecidos pelos pais, digitalizados e analisados pelo programa JAVA em textos com formato de arquivo XML. As listas geradas têm erros, ou *tokens*, com duas anotações cada: uma para a palavra correta e outra para o tipo de erro. A Tabela 3 ilustra um exemplo de resultado a partir da análise conduzida:

[4] O leitor é remetido ao texto original para detalhes sobre as categorias em árabe.

Tabela 3 – Frequência de Erros para os cinco tipos mais observados

Error word	Number of Occurences
On – علا	64
Which – اللتي	59
Which – اللذي	47
To – الى	35
That – ذالك	31

Fonte: Alamri e Thealan (2017)

Os erros mais observados foram adição de palavras (13.4%), omissão de palavra (10.98%), substituição (6.36%) e transposição (3.23%). Tais dados poderiam ser comparados a estudos feitos sobre dislexia com base em *corpora* em português, e fornecer um banco de dados apropriado é o objetivo principal da proposta do DYSCORP.

Entende-se que um *corpus* anotado possibilita a busca por erros, grupos ou tipos de erros específicos em termos de frequência de ocorrências. Observa-se certa dificuldade em encontrar estudos anotados sobre dislexia (Alamri; Teahan, 2017; Pedler, 2007; Rello, 2014): os poucos que existem, apresentam listas com erros frequentes e, em sua grande maioria, de fontes informais (Burhan *et al.*, 2014). No entanto, conforme apontado em Alamri e Thealan (2017) a partir de estudos com disléxicos em inglês, espanhol e dinamarquês, *corpora* com 1.000 palavras já parecem ser bastante úteis para propósitos de pesquisa.

4 AMBIENTE, APLICATIVOS E DISPOSITIVOS COM FOCO EM DISLEXIA

Ambientes, aplicativos e dispositivos desenvolvidos para o ensino de pessoas com dislexia são exemplos de inovação inclusiva. As iniciativas são interdisciplinares, resultado da articulação entre as Ciências da Computação e a Pedagogia, a Engenharia Elétrica e a Psicologia, a Psicopedagogia e o Design de Jogos e a Fonoaudiologia. Notamos, no entanto, que análises advindas da Linguística não figuram entre aquelas utilizadas para seu desenvolvimento.

Entre os recursos existentes, destacam-se ambientes virtuais, aplicativos e interfaces tangíveis (*tangible user interfaces* – TUIs – dispositivos que permitem a interação com dados digitais por meio de objetos físicos). Ambientes virtuais, como o *Mobile cloud computing* (Alghabban; Salama; Altalhi, 2017), combinam recursos audiovisuais e textuais, oferecendo lições e exercícios baseados em entrevistas com estudantes árabes disléxicos e em estudos bibliográficos sobre suas necessidades educacionais.

Da Colômbia, Chanchí, Sierra e Campo (2020) propõem um jogo educacional no qual um alienígena avança escolhendo palavras corretamente escritas, promovendo a internalização da ortografia do espanhol. O jogo foi desenvolvido a partir de questionários com usuários e da seleção de um conjunto de exercícios habituais utilizados nas terapias da dislexia. Outros jogos

para o espanhol partem das mesmas fontes, como o *Lee paso a paso*[5] e o *Palabras especiales*[6], que trabalham com a consciência fonológica de letras com traçados semelhantes (*n* e *u*, *m* e *w*), a contagem de sílabas e a relação entre palavra e imagem por meio de jogos de memória.

Em língua portuguesa, temos o *Mimosa e o reino das cores*[7], o *Arqueiro Defensor*[8] e o *Aramumo*. São jogos para crianças disléxicas que utilizam estratégias multissensoriais, como escutar e mover as sílabas para formar palavras ou escutar e mover palavras para construir frases. Já o *Desembaralhando* (Cidrim; Braga; Madeiro, 2018) é um aplicativo para intervenção no espelhamento ou inversão de letras (pares *b/d* e *a/e*) e na estruturação de frases. A seleção das palavras utilizadas no aplicativo parte de revisão da literatura na área e tem como base o *Alphabetics*[9], o *Hairry Letters*[10] e o *Dyseggxia*[11], que, por sua vez, tomam como ponto de partida o repertório de palavras mais frequentes na língua inglesa, as chamadas "sight words".

TUIs para o inglês, como o *PhonoBlocks* (Fan *et al.*, 2018), são abundantes. O *PhonoBlocks* utiliza letras 3D que mudam de cor para enfatizar mudanças sonoras. Os blocos foram construídos para crianças chinesas falantes de mandarim que aprendiam inglês como língua estrangeira. Para o espanhol, Ojeda-Zamalloa *et al.* (2020) desenvolvem o *Varinha Mágica*, que visa melhorar a coordenação motora em crianças de 4 a 6 anos alfabetizandas, disléxicas ou não. São três etapas: escrita com dedo, movimento de pinça em ferramenta 3D e a varinha mágica, que acompanha o movimento no traçado das letras. Para o português, o Alfaba (Jurgina, 2024) é um dispositivo multissensorial que combina estímulos visuais, táteis e auditivos, com letras tangíveis, luzes LED, saída de som e tela. O Alfaba faz seu usuário replicar palavras apresentadas de forma audível, dando reforço fonético e *feedback* visual. O dispositivo colore as letras conforme a situação: corretas brilham em verde, erradas em vermelho e espelhadas em azul.

É importante destacar que as iniciativas apresentadas são para crianças e nem todas são exclusivamente para crianças com dislexia. As iniciativas, em língua portuguesa e espanhola, desconsideram uma abordagem de *corpus*, especialmente no critério de seleção de palavras. Em língua inglesa, desde a década de 70, a constituição de listas de palavras como recurso educacional é discutida. A "sight word list" (Johns, 1972) é um exemplo disso: trata-se de uma lista de palavras de alta frequência em língua inglesa, constituída a partir da análise de quatro bancos de dados contendo palavras mais comuns em inglês. Essas palavras são popularmente chamadas de "palavras de reconhecimento visual" (*sight words*), as quais novos leitores devem memorizar visualmente e incorporar ao seu vocabulário visual logo no início do desenvolvimento da leitura.

Uma lista de palavras de reconhecimento visual em português, portanto, parece-nos premente para o desenvolvimento de recursos de apoio à dislexia. Nesse caminho, a *Dyslexic*

[5] https://apps.apple.com/us/app/lee-paso-a-paso-school-ed/id550555462?mt=8+lee

[6] https://itunes.apple.com/gt/app/palabras-especiales/id451723454?mt=8

[7] https://apkpure.com/br/mimosa-e-o-reino-das-cores/com.itabits.colorrindo

[8] http://arqueiro-defensor.apk.watch/1.0.1b

[9] https://itunes.apple.com/br/app/alphabetics/id621297604?mt=8 alphabetics

[10] https://itunes.apple.com/br/app/hairy-letters/id997423057?mt=8

[11] http://dyslexiahelp.umich.edu/tools/apps

Sight Words para o português é proposta por Cidrim, Azevedo e Madeiro (2021) a partir de questionário aplicado com educadores em contato com crianças disléxicas. Esses profissionais constituíram uma lista de 40 palavras que escolares com dislexia frequentemente escrevem de modo incorreto. Sabe-se que os erros são naturais no processo de apropriação do sistema ortográfico, desvelando as etapas do domínio da consciência ortográfica, mas erros específicos e assistemáticos podem auxiliar a predizer dislexia. Resta, no entanto, determinar a especificidade e a assistematicidade dos erros manifestados por pessoas com dislexia por meio de *corpora* e não apenas através da aplicação de questionários, metodologia vigente, conforme vimos no ambiente, aplicativos, dispositivos apresentados e na própria *Dyslexic Sight Words* para o português.

A Linguística de *Corpus* e a Linguística Computacional, nesse sentido, muito têm a contribuir nessa direção. Por exemplo, Oliveira e Capellini (2016) compilaram um *corpus* a partir de materiais didáticos de língua portuguesa utilizados pela rede de ensino do Estado de São Paulo. Com isso, apresentam um banco de palavras para leitura de escolares do Fundamental II, uma pista para a lista de sight words em língua portuguesa, restando compilar para Fundamental I e Educação Infantil. Gazzola, Leal e Aluisio (2019) compilaram um grande *corpus* com textos utilizados em diferentes etapas de ensino do Sistema Educacional Brasileiro. O *corpus* foi organizado em quatro etapas utilizadas na Plataforma MEC de Recursos Educacionais Digitais (MEC-RED) para classificação nos estágios escolares: Ensino Fundamental I (primeiro ao quinto ano); Ensino Fundamental II (sexto ao nono ano); Ensino Médio; Ensino Superior. O *corpus* está publicamente disponível e inclui livros-texto, notícias da seção Para Seu Filho Ler (PSFL) do jornal Zero Hora, Exames do SAEB, Livros Digitais do Wikilivros em português e Exames do Enem.

Com base nas ferramentas recém ilustradas e nos exemplos de corpora já existentes sobre dislexia abordados na seção anterior, reforçamos o caráter inovador da proposta do DYSCORP, que deverá permitir acesso a informações mais específicas sobre dislexia em níveis lexicais e sintáticos através de conteúdo digitalizado organizado para fins de investigações. Os dados serão extraídos de material utilizado no Projeto ACERTA[12] (Avaliação de Crianças em Risco de Transtorno de Aprendizagem), do Instituto do Cérebro (InsCer), já disponível e utilizado pelo grupo de pesquisa *NeuroEdTech Lab da PUCRS.* O material é constituído de textos e listas de palavras produzidos por crianças disléxicas e foi coletado para atender a propósitos de pesquisas anteriores (Azevedo, 2016), durante o período de 2013-2019. O projeto ACERTA surgiu com o propósito de investigar crianças em fase de alfabetização a fim de identificar preditores de dificuldades de aprendizagem. O ACERTA teve como objetivo compreender os mecanismos neurais associados com a alfabetização e os obstáculos que a eles se impõem nos transtornos de aprendizagem. Em conjunto com outros dois centros de pesquisa, em Florianópolis e Natal, o projeto buscou identificar biomarcadores precoces desses transtornos, através do uso da neuroimagem funcional e estrutural. Paralelamente, o projeto visou divulgar e conscientizar a comunidade escolar sobre os transtornos específicos de leitura (dislexia) e matemática (discalculia), que afetam entre 5 e 10% das crianças em todo o mundo.

[12] https://inscer.pucrs.br/br/noticias/interna/projeto-acerta.

Tal como proposto em Pacheco (2024), as etapas cobrem desde a digitalização do material disponível organizado de forma a seguir os princípios da Linguística de *Corpus* até a criação de taxonomias específicas para análise dos dados. O trabalho é realizado em âmbito de projeto de pesquisa acadêmico e, por isso, conta com a participação de uma equipe multidisciplinar como linguistas, neurocientistas, fonoaudiólogos, psicólogos e pedagogos.

5 A PROPOSTA DYSCORP

O DYSCORP tem sido compilado, inicialmente, como parte de um projeto acadêmico da Pontifícia Universidade Católica do Rio Grande do Sul vinculado ao grupo de pesquisa NeuroEdTechLab da Escola de Humanidades. As informações estão sendo extraídas de dados previamente utilizados para fins do Projeto ACERTA, que foram devidamente organizados em planilhas de Excel por membros do grupo de pesquisa *NeuroEdTechLab* para fins específicos do Projeto DYSCORP, tal como mostra a Figura 2 abaixo:

Figura 2 – Extrato de planilha utilizado na compilação do DYSCORP
Fonte: Autoras (2025)

	B	C	D	E	F	G	H	I	J
			PALAVRA						
	SUJEITO	IDADE	sapo	bola	zure	alimento	cratilo	conversa	noite
	88	10	Saber	Bela	Zirre	Alimejo	Carajulo	Conveza	Neco
	89	9	sapo	bala	zura	alimento	cratilo	conversa	noite
	90	10	sapo	bola	zurre	alimento	cratrilo	conversa	noite
	91	12	sapo	bola	zure	alimento	cratilo	conversa	noite
	92	9	sapo	bola	surra	alimento	cartivo	conversa	noite
	93	12	sapo	bola	zure	alimento	cretilo	conversa	noite
	94	9	sapo	bola	surre	alimento	cratilho	conversar	noite
	95	9	sapo	bola	vurre	x	cerrapilo	x	tunite
	96	11	sapo	bola	zure	alimento	cratilo	conversa	noite
	97	11	sapo	bola	zure	alimento	cratilo	conversa	noite
	98	11	sapo	bola	zuer	alimento	cartilo	conversa	noite
	99	10	sapo	bola	zure	alimento	corto	conversa	nonte
	100	9	sapo	bola	zore	alimento	cratilo	conversa	noite
	101	11	sapo	bola	zurre	alimento	cratilo	conversa	noite
	102	10	sapo	bola	zurre	x	x	x	x
	103	10	sapo	bola	zurre	alimento	cranti	conversar	muinto

Dessas informações, organizadas a partir de material do projeto, está sendo compilado um *corpus* – um banco de dados organizado de forma a poder ser explorado por ferramentas especializadas para fins de investigação linguística. Em estágio preliminar, o *corpus* conta com 2176 palavras, organizadas provisoriamente em Documento Word, seguindo a orientação:

XX <PO_XX.PR_XX.SU_XX.ID_XX>

Sendo, nesse exemplo, as letras apresentadas indicações de:

XX: informação a ser preenchida no *corpus* a partir da coleta

PO: palavra original, ou seja, a palavra-alvo oriunda de dados utilizados no Projeto Acerta.

PR: protocolo, igualmente a partir de referência do Projeto Acerta

SU: sujeito, identificados por números.

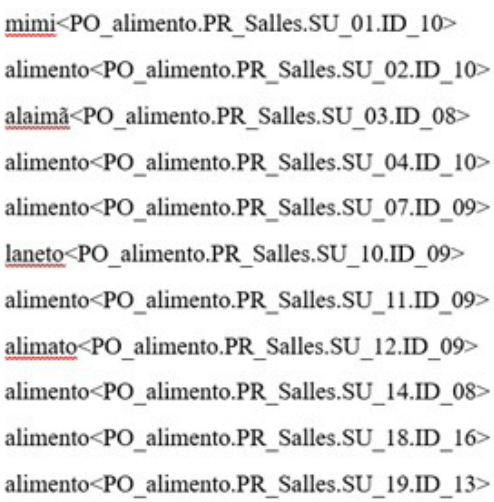
ID: idade, identificada também por números.

Com base nessas orientações, temos abaixo um exemplo do *corpus*

sapo <PO_sapo.PR_Salles.SU_04.ID_10>

Aqui, a palavra escrita pelo sujeito foi "sapo", da palavra original "sapo", do Protocolo Salles, do sujeito 04, com idade de 10 anos. A Figura 3 abaixo ilustra uma parte do arquivo.

Figura 3 – Ilustração Preliminar do DYSCORP



Fonte: Autoras (2025)

Num primeiro momento de compilação, para fins desse projeto acadêmico, serão compilados dados oriundos de 108 palavras coletadas de 103 sujeitos de dois protocolos de leitura e escrita, a saber, Avaliação da leitura de palavras e pseudopalavras (Salles, 2005) e Ditado balanceado (Moojen, 2009). Num segundo momento, pretende-se inserir categorias de etiquetagem em cada linha, adicionalmente à informação já disponibilizada entre os sinais < e >. Ou seja, será inserida outra informação referente à categorização da produção do sujeito. Como posto na seção 3, corpora associados à dislexia são escassos na literatura e categorias de etiquetagem parecem ser bastante peculiares aos propósitos dos estudos. No caso do projeto desenvolvido, as seguintes categorias expostas na Tabela 4 estão sendo pensadas como possíveis marcadores:

Tabela 4 – Sugestões de Categorização de Erros para o DYSCORP

PC	Palavra correta	Palavra correta
LE	Letras espelhadas	Inversão de "b" por "d" ou "p" por "q", por exemplo, relacionada a dificuldades de orientação espacial e memória visual
SB	Substituição	Troca de letras com sons ou formas gráficas semelhantes (por exemplo, escrever "faca" como "vaca")
IN	Inserção	Letras ou sílabas são adicionadas indevidamente a uma palavra (como "escreber" em vez de "escrever")
AP	Apagamento	Apagamentos ou omissões, que consistem na exclusão de letras ou sílabas essenciais (por exemplo, "esoa" em lugar de "escola")
TR	Transposição	Troca na ordem das letras ou sílabas, como em "porblema" em vez de "problema"

Fonte: Autoras (2025)

A partir disso, o *corpus* poderá ser explorado por ferramentas específicas para investigação linguística a fim de fornecer informações referentes à frequência de ocorrências – da palavra correta ou do tipo de erro, por exemplo, ou de produções de sujeitos com uma determinada faixa etária.

O DYSCORP pretende fornecer uma base de dados sólida, robusta, organizada e acessível que poderá ser usada por educadores, linguistas, psicólogos e outros profissionais que tenham interesse na pesquisa com o português brasileiro e sobre a dislexia. Esses profissionais poderão identificar padrões de erros característicos da dislexia por meio dos recursos produzidos nesta pesquisa. O *corpus* terá princípio de total abertura, sendo público e com todos os seus recursos abertos tão logo seja licenciado pelos proponentes. Seu princípio é criar condições para a existência de recursos gratuitos para o estudo da dislexia em português brasileiro e outras línguas, ou seja, a sua disponibilização será um resultado mais imediato a partir do início da pesquisa.

Vislumbra-se a criação de novos produtos tecnológicos e educacionais que possam ser disponibilizados gratuitamente, com o intuito de inovar a educação brasileira e, sobretudo, a inclusão. A prototipagem de um aplicativo baseado no *corpus* proporcionará ferramentas práticas para diagnóstico precoce e acompanhamento de indivíduos com dislexia. O *app* deverá incluirá funcionalidades como análise automática de erros e sugestões adaptadas para aprimorar a leitura e a escrita.

6 CONSIDERAÇÕES FINAIS

Este artigo apresentou a proposta do DYSCORP, um banco de dados oriundos de sujeitos disléxicos organizado de maneira a ser explorado por ferramentas específicas à luz da Linguística de *Corpus*. Em todas as suas instâncias – teórico-metodológica, resultados e divulgação – a proposta deverá gerar impactos para diferentes esferas sociais dado o seu caráter inclusivo e inovador. Por meio da pesquisa realizada, prevê-se que as contribuições oriundas das reflexões teóricas e artigos produzidos servirão para desmistificar informações correntes a respeito da dislexia. A ciência da leitura e o *corpus* a ser compilado e entregue à sociedade científica e não especializada apontará caminhos para abordar estrategicamente a dislexia nos contextos educacional e social brasileiros.

À medida que indivíduos disléxicos têm suas dificuldades mais bem compreendidas, abre-se um outro espectro de possibilidades para que possam ter suas necessidades igualmente mais bem atendidas. No âmbito do ensino, professores com mais informação sobre tipos de desvios linguísticos frequentemente evidenciados por tais sujeitos poderão não somente preparar materiais que incluam esses possíveis problemas como também gerenciar a ocorrência de problemas na esfera da sala de aula de modo a mitigar quaisquer danos que a exposição da divergência possa acarretar.

No âmbito do tratamento especializado, profissionais de saúde com acesso a informações referentes a tipos e taxonomias de erros poderão criar protocolos para atendimento mais padronizados que possam endereçar a administração desses desvios de modo mais específico. Este artigo limita-se a apresentar o DYSCORP enquanto proposta de pesquisa em estágio preliminar de execução. Espera-se que artigos adicionais possam explorar em mais detalhes tanto subsídios teóricos que embasam a proposta quanto descrições dos dados investigados e dos recursos tecnológicos efetivamente derivados do projeto, a fim de que a dislexia possa ser mais bem compreendida e tratada.

REFERÊNCIAS

ALAMRI, M.; TEAHAN, W. A New Error Annotation for Dyslexic texts in Arabic. *Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP)*. Valencia: Association for Computational Linguistics, 2017.

ALGHABBAN, W. G.; SALAMA, R. M.; ALTALHI, A. H. Mobile cloud computing: An effective multimodal interface tool for students with dyslexia. *Computers in Human Behavior*, Amsterdam, v.75, p. 160-166, 2017.

ASSOCIAÇÃO NACIONAL DE DISLEXIA – AND. *Perguntas Frequentes*. *[S. l.: s. n.]*. [20--?]. Disponível em: https://www.andislexia.org.br/. Acesso em: 23 jan. 2024.

AZEVEDO, A. *Cérebro, leitura e dislexia*: um estudo experimental sobre a leitura e as bases neurais da dislexia em monolíngues e aprendizes de inglês como L2, com o uso de ressonância magnética funcional. Orientador: Augusto Buchweitz. 2016. Tese (Doutorado em Letras) – Pontifícia Universidade Católica do Rio Grande do Sul, Faculdade de Letras, Porto Alegre, 2016. Disponível em: https://hdl.handle.net/10923/9623. Acesso em: 23 jan. 2024.

AZEVEDO, A. *et al.* Dyslexia and the perks of being bilingual: A study on the neurobiology of reading with fMRI. *Ilha Do Desterro*, [*s. l.*], v. 76, n. 3, p. 279-299, 2023. Disponível em: https://doi.org/10.5007/2175-8026.2023.e94524. Acesso em: 23 jan. 2024.

BERBER SARDINHA, T. *Linguística de Corpus*. São Paulo: Manole, 2004.

BIBER, D; REPPEN, R.; FRIGINAL, E. Research in *Corpus* Linguistics. *In:* KAPLAN, R.B. (ed.). *The Oxford Handbook of Applied Linguistics.* 2. ed. Oxford: Oxford University Press, 2010. p. 548-570.

BURHAN, H. *et al.* Degree of Common Misspellings of Students with Learning Disabilities. *Interdisciplinary Journal of Education*, [*s. l.*], v. 3, p. 1-11, 2014.

CHANCHÍ, G. E. G.; SIERRA, L. M. M.; CAMPO, W. Y. M. Propuesta de un videojuego educativo como apoyo a las terapias de la dislexia, usando la plataforma GDevelop. *Revista Ibérica de Sistemas*

e Tecnologias de Informação, Portugal, n. E29, p. 173-86, 2020.

CIDRIM, L.; AZEVEDO, N.; MADEIRO, F. Elaboração de uma lista de palavras no âmbito da ortografia para escolares com dislexia: 'Dyslexic Sight Words'. *Revista Psicopedagogia*, São Paulo, v. 38, n. 115, p. 5-17, 2021.

CIDRIM, L.; BRAGA, P. H. M.; MADEIRO, F. Desembaralhando: um aplicativo para a intervenção no problema do espelhamento de letras por crianças disléxicas. *Revista CEFAC*, Campinas, v. 20, p. 13-20, 2018.

EVERS, A. *A redação engaiolada*: padrões lexicais e ensino de redação em cursos pré-vestibulares populares. Orientador: Maria José Bocorny Finatto. 2018. 297 f. Tese (Doutorado em Letras) – Universidade Federal do Rio Grande do Sul, Instituto de Letras, Porto Alegre, 2018.

EVERS, A. *GALAX(IA):* sistema de retroalimentação para textos escritos. Porto Alegre: Projeto Edital de Professor Horista da Pontifícia Universidade Católica do Rio Grande do Sul, 2024.

EVERS, A. *Processamento de língua natural e níveis de proficiência do português: um estudo de produções textuais do exame Celpe-Bras*. 2013. Dissertação (Mestrado em Estudos da Linguagem) – Programa de Pós-Graduação em Letras, Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013.

FAN, M. *et al.* Block talks: A tangible and augmented reality toolkit for children to learn sentence construction. *In:* Extended Abstracts of the 2018 Chi Conference on Human Factors in Computing Systems, 2018. *Anais* [...]. [*S. l.: s. n.*], 2018. p. 1-6.

FREITAS, C. Estudos linguísticos e Humanidades digitais: *corpus* e descorporificação. *Gragoatá*, [*S. l.*], v. 22, n. 44, p. 1207-1227, 2017.

GAZZOLA, M. G.; LEAL, S. E.; ALUÍSIO, S. M. Predição da complexidade textual de recursos educacionais abertos em português. 2019, Porto Alegre. *Anais* [...]. Porto Alegre: SBC, 2019.

GRANGER, S. A. Bird's eye view of learner *corpus* research. *In:* GRANGER, S.; HUNG, J.; TYSIN, P. S. (ed.). *Computer learner corpora, second language acquisition and foreign language teaching.* Amsterdam: John Benjamins Publishing Company, 2002.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. Censo Demográfico 2010. Rio de Janeiro: IBGE, 2012.

JOHNS, J. L. A List of Basic Sight Words for Older Disabled Readers. *The English Journal*, v. 61, n. 7, p. 1057-1059, 1972.

JURGINA, L. Q. *Alfaba [recurso eletrônico]:* desenvolvimento de uma ferramenta multissensorial com interface tangível para o suporte à alfabetização de crianças com dislexia. Orientador: Tiago Thompsen Primo. Coorientador: Marilton Sanchotene de Aguiar. 2024. 95 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2024.

McENERY, T.; WILSON. A. *Corpus linguistics*. Edinburgh: Edinburgh University Press, 1996.

MOOJEN, S. *A escrita ortográfica na escola e na clínica:* Teoria, avaliação e tratamento. São Paulo: Casa do Psicólogo, 2009.

OJEDA-ZAMALLOA, I. *et al.* Diseño, construcción e implementación de una herra-mienta para el soporte para ejercicios de grafo-motricidad en niños de cuatro a seis años. *Revista Ibérica de Sistemas e Tecnologias de Informação*, Portugal, n. E33, p. 263-274, 2020.

OLIVEIRA, A. M. DE .; CAPELLINI, S. A. E-LEITURA II: banco de palavras para leitura de escolares do Ensino Fundamental II. *CoDAS*, v. 28, n. 6, p. 778-817, 2016.

PACHECO, A. *A aquisição de morfemas em inglês como L2*: uma análise dos padrões evolutivos através do BELC (*Brazilian English Learner Corpus*). 2010. 188 f. Tese (Doutorado em Letras) – Universidade Federal do Rio Grande do Sul, Instituto de Letras, Porto Alegre, 2010.

PACHECO, A. *DYSCORP (**Dyslexia** Corpus):* compilação de um *corpus* para propósitos específicos no estudo da dislexia. Porto Alegre: Projeto Edital de Professor Horista da Pontifícia Universidade Católica do Rio Grande do Sul, 2024.

PACHECO, A. *et al.* Using *corpus* linguistics to create tasks for teaching and assessing Aeronautical English. *Applied Corpus Linguistics*, [*s. l.*], v. 3, n. 3, 2023.

PEDERSEN, J.; LARSEN, L. A Speech *Corpus* for Dyslexic Reading Training. *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC'10). Valletta: European Language Resources Association (ELRA), 2010.

PEDLER, J. *Computer Correction of Real-word Spelling Errors in Dyslexic Text*. 2007. Tese (Doutorado) – Birkbeck College, London University, London, 2007.

PEDLER, J.; MITTON, R. A Large List of Confusion Sets for Spellchecking Assessed Against a *Corpus* of Real-word Errors. *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC'10). Valletta: European Language Resources Association (ELRA), 2010.

RELLO, L. Design of Word Exercises for Children with Dyslexia. *Procedia Computer Science*, [*s. l.*], v. 27, p. 74-83, 2014.

RELLO, L.; BAEZA-YATES, R.; LLISTERRI, J. DysList: An annotated resource of dyslexic errors. *LREC 2014 Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavik: European Language Resources Association (ELRA), 2014.

RODRIGUEZ-SEGURA, D. EdTech in Developing Countries: A Review of the Evidence. *The World Bank Research Observer,* v. 37, n. 2, p. 171-203, 2022.

SALLES, J. F. *Habilidades e dificuldades de leitura e escrita em crianças de 2ª série:* abordagem neuropsicológica cognitiva. Tese (Doutorado) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2005.

TAGNIN, S. E. O. *O jeito que a gente diz*: as expressões convencionais e idiomáticas. Ed. rev. e ampl. São Paulo: Disal, 2013.