Relatório de estágio pós-doutoral

Scriba - uma IA aplicada ao ensino de ortografia

Adelaide H. P. Silva (DELLIN/UFPR; PGInf/UFPR)
Supervisor: prof.dr. Fabiano Silva (DInf/PGInf/UFPR)

1 - Introdução

Este projeto de pós-doutorado tinha inicialmente o título "Subsídios para um corretor ortográfico multilingue", mas ele sofreu alterações até chegar no formato final que será apresentado neste relatório. O estágio pós-doutoral, que se circunscreve à área de Processamento de Linguagem Natural, foi desenvolvido no período de 01 de março de 2020 a 28 de fevereiro de 2021 no Programa de Pós-Graduação em Informática da UFPR, sob a supervisão do prof. dr. Fabiano Silva.

Antes do início oficial do estágio, eu já frequentava o Laboratório de Inteligência Artificial e Métodos Formais (LIAMF), coordenado pelo prof. Fabiano. Isso me possibilitou contato com os alunos de doutorado, mestrado e Iniciação Científica que frequentam o Laboratório e, como consequência, já no período oficial do pós-doc pude participar de reuniões e me inteirar dos projetos ali desenvolvidos. A proposta do prof. Fabiano, inclusive, era a de que eu co-orientasse alguns trabalhos, como de fato ocorreu e conforme eu comento mais adiante neste relatório.

Entretanto, a pandemia do novo coronavírus promoveu mudanças de alguns planos inicialmente previstos. Um desses planos foi o objeto do próprio projeto: inicialmente, nosso objetivo era desenvolver um corretor ortográfico que funcionasse em tempo real e que conseguisse processar várias línguas. No limite, como eu afirmava no texto do projeto, essa ferramenta poderia ser até mesmo empregada como recurso adicional em ensino de português para estrangeiros.

Com o distanciamento social inevitável, o fechamento das escolas e a consequente adoção de atividades remotas de ensino, surgiu-me a ideia de reformular o projeto, de modo a propor uma ferramenta que pudesse auxiliar professores e alunos no processo de ensino/aprendizagem e que pudesse ser utilizado tanto em atividades presenciais como em atividades remotas.

Inicialmente, cogitei um aplicativo que oferecesse um jogo para trabalhar ortografia, a exemplo do que existe para outras línguas como, e.g., o espanhol (Ortografía española, desenvolvido por Eductify[1]). Por essa razão, o prof. Fabiano me colocou em contato com o prof. dr. Roberto Pereira,

1 Jogo multiplataforma e com versões para Android e IOS, voltado para crianças e adultos e, aparentemente, bem aceito pelo público. Pode ser encontrado em https://www.eductify.com/es/ortografia-espanola.

especialista em Interação Humano-Computador. Eu cheguei a programar um jogo, considerando os requisitos básicos que o sistema demandaria. Esse era um momento ainda muito inicial de todo o processo, porque elaborar um jogo requereria muito mais do que desenvolver a programação de um. Era preciso, e.g., pensar a que público o jogo se destinaria, para então construir uma interface minimamente atraente aos jogadores[2]. A conversa com ambos os professores direcionou o projeto para um novo caminho e, ainda considerando a ortografia, o prof. Fabiano e eu decidimos por construir uma inteligência artificial (IA) que simulasse "erros" de crianças em fase de alfabetização. Tal decisão tinha por objetivos verificar se seria possível reconhecer um padrão nos "erros" e associá-los ao nível de instrução formal das crianças. Como decorrência, previmos a possibilidade de oferecer uma ferramenta que permitisse aos professores avaliar os alunos em função desses padrões e, adicionalmente, oferecer insumos para diretrizes ao ensino de ortografia, como deverá ficar mais claro mais adiante neste texto.

Esse foi o rumo que o projeto tomou e, atualmente, a IA já passou pela fase inicial de aprendizado e consegue clusterizar - ou agrupar - alguns conjuntos de "erros" por semelhanças. São necessárias melhorias no agrupamento, e a discussão atual concerne à verificação do algoritmo que tenha melhor desempenho nessa tarefa.

Em sua sequência, este relatório esclarece por que a opção pela ortografia, bem como apresenta de forma resumida os passos que se seguiram para a formulação da IA. É preciso frisar, desde este momento, que apesar de o estágio pós-doutoral ter se encerrado em 28/02, o projeto deve continuar, assim como a parceria iniciada no Departamento de Informática (DInf) a partir do meu pós-doutorado. Essa, aliás, foi a grande vantagem de eu ter podido realizar o estágio na própria UFPR: o pós-doc não se encerra com o final de meu afastamento formal; ao contrário, inicia uma parceria promissora e que visa a integrar duas áreas do conhecimento, produzindo ciência, tecnologia e inovação.

1 - A opção pela ortografia

De acordo com Chacon e Pezarini (2018), "a escrita é um modo de enunciação da língua, ou seja, uma forma de colocar a língua em uso em uma dada situação discursiva, para que se possa aprender a escrever, de acordo com Morais e Leite (2012), bem como de acordo com os Parâmetros Curriculares Nacionais da Língua Portuguesa (1997) – PCNs –, é necessário compreender tanto a

2 Quando apresentei meu projeto nos Seminários da Pós-Graduação em Informática, em 05 de junho de 2020, a ideia ainda era elaborar um jogo que pudesse auxiliar seus usuários a aprender ortografia de maneira contextualizada, ou seja, incluindo palavras-alvo em contextos de uma narrativa, por exemplo. Naquele momento, o público-alvo eram crianças do Ensino Fundamental II (sexto ao nono ano). Mas, como eu exponho na sequência, esse plano sofreu algumas alterações.

natureza do sistema de escrita da língua (o sistema notacional), quanto o funcionamento de (seus) aspectos discursivos (o uso social da escrita)."

Como os autores, parto da premissa de que a ortografia - o sistema notacional de uma língua - deve ter seu funcionamento compreendido pelos usuários para que eles possa aprender a escrever e, assim, "colocar a língua em uso em uma dada situação discursiva". Ou seja, aprender a ortografia de uma língua se configura como o primeiro passo dos usuários dessa língua rumo ao aprendizado da escrita. Não é à toa que o sistema notacional de uma língua é abordado nos anos iniciais de instrução formal dos indivíduos.

Apesar de ser abordada nos anos escolares iniciais, o ensino da ortografia não é sistematizado, i.e., não se prevê o que e quando ensinar, como se o sistema notacional da língua se resumisse à correspondência entre som e letra e tal correspondência fosse biunívoca, com algumas poucas exceções. Entretanto, Lemle (1982) já elencava três estágios que, segundo ela, norteariam o processo de alfabetização:

1) biunivocidade, que consiste no fato de que que cada grafema representa um som da fala e cada som é representado por um grafema, e.g., <f> representa [f]; <p> representa [p]; <b> representa [b][3];

2) biunivocidade contextualizada, que implica que um grafema representa um som da fala, num contexto específico e cada som é representado por um grafema num determinado contexto, como "genrro" por "genro" (aprendiz sabe que <rr> representa o "r forte" entre vogais, no meio da palavra, como em "erro", e generaliza para outros contextos); ou "tenpo" por "tempo" (aprendiz sabe que <n>, depois de vogal, representa a nasalidade da vogal, mas ainda não aprendeu que antes de <p,b> a nasalidade da vogal é sinalizada por <m>;

3) equivalência entre letras e idiossincrasias ortográficas, que implica que grafemas representam sons da fala e, na maioria dos casos, cada som, num contexto específico, é representado por um grafema. Mas alguns sons, num determinado contexto, admitem mais de uma representação e, nesses casos, a representação precisa ser memorizada, como o som [ʃ], que pode ser grafado com <x>, em "xale" ou <ch> em "chalé". Segundo Lemle (*op.cit.*), ao atingir o terceiro estágio, o indivíduo pode ser considerado alfabetizado.

Chacon e Pezarini (*op.cit.*) apresentam uma releitura desses estágios, propondo que o processo de alfabetização envolva correspondências transparentes entre grafemas e sons, as quais corresponderiam ao estágio (1) previsto por Lemle, além de correspondências opacas entre grafemas e sons e que corresponderiam aos estágios (2) e (3) preconizados por Lemle. Essa proposta de Chacon e Pezarini (2018) leva-os a observarem que correspondências opacas entre

3 Entre < > dispõem-se os grafemas, ou letras; entre [ ], os sons da fala. Esta observação se aplica a todo o texto.

grafemas e sons são regidas por convenções, que determinam o uso de um dado grafema para representar um som, considerado ou não o contexto em que tal som ocorre. Como decorrência das correspondências opacas, as convenções organizam e normatizam a ortografia, estabelecendo uma forma de grafia das palavras que independe de possíveis variações fonológicas.

Assim, Chacon e Pezarini (2018) argumentam que o aprendizado da ortografia envolve não apenas a compreensão sobre a relação entre sons e grafemas, mas também sobre a convenção que permeia, de modo mais transparente ou mais opaco, a relação entre essas unidades.

Desse modo, ao mesmo tempo em que notam que o acertar a grafia de uma palavra não implica, necessariamente que uma criança conheça as características do sistema ortográfico da língua portuguesa, os "erros" não são necessariamente um sinal de que a criança desconheça tais características. Antes, podem traduzir hipóteses que a criança faz sobre a organização da ortografia da língua.

Neste ponto, cabem parênteses para observar que, ao aprender os "erros" que as crianças em fase de alfabetização produzem e agrupá-los em função de semelhanças entre eles - e.g. "esplodiu", "ispludiu", "ispludio" para "explodiu" - nossa IA pode oferecer pistas importantes sobre hipóteses subjacentes ao processo e que levem a uma melhor compreensão dos "erros".

Voltando ao artigo de Chacon e Pezarini (*op. cit.*), os autores concebem que haja uma gradiência nas relações entre os aspectos fônicos do português brasileiro e o sistema ortográfico. Esta é uma nova visão na literatura, que difere de abordagens anteriores (e.g., LEMLE, 1982; CAGLIARI, 1989; MORAIS, 2000), as quais preveem categorias de "erros". Chacon e Pezarini (2018), por sua vez, preconizam que haja processos de omissões, transposições e substituições. No primeiro caso, teríamos, e.g., "pene" por "pente"; no segundo, "porfessor" por "professor" e, no terceiro, "sebola" por "cebola". Os autores explicam que a gradiência se revela no fato de que, nas omissões, inexiste o registro ortográfico de um som, o que as coloca numa situação distinta daquela de transposições e substituições, nas quais existe o registro ortográfico do som que se deseja representar, ainda que nas transposições o grafema que representa um determinado som não se encontre na posição esperada e, nas substituições, o grafema ocupe a posição esperada, mas seja, ele mesmo, esperado na representação de um determinado som.

Ainda no que concerne a transposições e substituições, os autores argumentam que há gradiência no interior de cada fato. Assim, transposições podem envolver permutas, quando há troca de posição de dois grafemas no interior de uma palavra (e.g., "senera" por "serena"), ou quando há troca de grafemas intersílabas (e.g, "drento" por "dentro", em que a troca de posição de um grafema afeta duas sílabas) ou, ainda quando há troca de grafemas intrassílabas e um grafema se desloca de uma posição para outra na mesma sílaba (e.g, "pregunta" por "pergunta", em que a troca de posição

de um grafema afeta uma única sílaba). No que diz respeito às substituições, Chacon e Pezarini (2018) notam que elas podem ser: substituições híbridas (nos casos em que um grafema é substituído por outro que pode representar o mesmo fonema, noutro contexto, como "lícido" por "líquido"); substituições não-fonológicas (nos casos em que um grafema é usado para representar um som num contexto em que não era previsto, como "rrato" por "rato"); substituições ortográficas fonológicas (nos casos em que a alteração de um grafema altera o que os autores denominam "valor fonológico", como em "galo" por "calo", em que se altera a sonoridade da consoante representada pelo grafema inicial da palavra). Cabe comentar que os autores acrescentam que as substituições ortográficas fonológicas se comportam de modo distinto das outras substituições e podem envolver sons de uma mesma classe, bem como sons de classe diferente. Assim, e.g., uma substituição de grafema que representa som de classe distinta do grafema-alvo pode ser encontrada em "molacha" por "bolacha". (Ainda assim, parece-me necessário comentar que há uma grande semelhança articulatória entre [b] e [m], já que ambos são articulados no mesmo ponto do trato vocal, envolvem oclusão labial e requerem vibração das pregas vocais. A única diferença é que [m] requer, adicionalmente, o escape de fluxo de ar pela cavidade nasal, ao contrário de [b]. Ou seja, as substituições não são aleatórias.) Já as substituições de grafemas que representam sons dentro de uma mesma classe podem ser exemplificadas pela grafia "gola" por "cola".

É necessário considerar que a proposta de Chacon e Pezarini (2018) é um trabalho em andamento, por isso não contempla "erros" envolvendo grupos consonantais ou consoantes em coda silábica (posição final de sílaba).

Ainda assim, a proposta incorpora informações fonológicas que devem permitir postular uma motivação para os "erros" envolvendo relações menos opacas entre grafemas sons. Como consequência, podemos prever dois desdobramentos: 1) a IA que estamos construindo pode testar a proposta de Chacon e Pezarini (2018), já que a clusterização dos "erros" pode revelar se a informação fonológia é um bom parâmetro para esse fim e se há gradiência na produção dos "erros"; 2) o agrupamento dos "erros" em *clusters* e a observação dos parâmetros que motivam a clusterização podem oferecer subsídios para o ensino de ortografia.

Cabem, adicionalmente, nesta seção, algumas considerações sobre o ensino de ortografia, e que em boa medida justificam a preocupação por fornecer subsídios a ele, através da construção da IA: Morais (2000: 66) nota que "... como a ortografia é tratada entre nós mais como tema de verificação que de ensino sistemático, a maioria das escolas do país funciona sem planejar o que espera conseguir na promoção da competência ortográfica de seus alunos a cada série. E como quem não tem metas não antevê aonde quer chegar, não planifica sua ação... pode não conseguir progressos significativas no rendimento que seus alunos expressam ao escrever."

A falta de planejamento identificada por Morais (*op.cit.*) se reflete na ausência de diretrizes para o ensino/aprendizagem de ortografia no texto da Base Nacional Curricular Comum (BNCC). Esse documento dispõe, em sua versão de 2018, no que concerne às "competências específicas de linguagens para o ensino fundamental" e, mais especificamente, para os anos iniciais dessa etapa de escolarização, que uma das habilidades a ser desenvolvida é a "fono-ortografia". Essa habilidade consistiria em "Conhecer e analisar as relações regulares e irregulares entre fonemas e grafemas na escrita do português do Brasil. Conhecer e analisar as possibilidades de estruturação da sílaba na escrita do português do Brasil." (BNCC, 2018:82) Essa é a única menção à aprendizagem de ortografia em todo o texto da BNCC, apesar de o domínio da ortografia ser parâmetro para avaliar se uma criança está ou não alfabetizada, ao final do primeiro ciclo do Ensino Fundamental I (terceiro ano). Resultado: em sua última edição, datada de 2016, a Avaliação Nacional de Alfabetização (ANA) acusava índices baixos na avaliação da escrita: segundo informações do próprio MEC, 33,95% dos estudantes ainda exibiriam nível insuficiente em seu desempenho nesse quesito[4]. A necessidade de se pensar numa ferramenta que possa auxiliar o processo de alfabetização fica, portanto, clara e óbvia. E é urgente.

2 - Elaboração da Inteligência Artificial

2.1 - *Intelligent Tutoring Systems* (ITS)

A IA que estamos construindo segue os princípios dos Sistemas Tutores Inteligentes, ou *Intelligent Tutoring Systems* (ITS). A sugestão do prof. Fabiano de enveredarmos por essa perspectiva resulta do esforço dos colegas do Departmento de Informática em elaborar sistemas que possam, de alguma forma, oferecer soluções para questões que envolvam o ensino em geral e que orienta muitos dos sistemas desenvolvido no Centro de Computação Científica e Software Livre (C3SL) .

Os ITS se constituem de dois componentes: um modelo de execução, capaz de realizar as tarefas que os alunos estão aprendendo, e um modelo de ensino, que compara as ações dos alunos com aquelas engendradas pelo modelo de execução e determina como o ITS responde ao aluno.

Nossa IA não é propriamente um ITS, embora se baseie nos princípios de um, como comentei anteriormente. Como um ITS, ela contém um modelo de execução, constituído de um conjunto de regras inferenciais. Tal modelo é a base da nossa IA, pois ela deve conseguir fazer generalizações sobre os "erros" ortográficos, não só de forma a produzir "erros" que alunos de ensino fundamental produziriam, mas também de modo a agrupá-los por semelhanças. Esse procedimento deverá

4 Informações disponíveis em http://basenacionalcomum.mec.gov.br/abase/#fundamental/lingua-portuguesa. Acesso em 02/03/2021.

possibilitar, adicionalmente, que se associem "tipos de erros" a um momento da instrução formal dos alunos no qual tais erros são mais esperados.

O modelo de execução da IA deverá possibilitar uma espécie de modelo de ensino: comparando as produções de um determinado aluno com as produções que podem ser geradas pelo modelo de execução, a IA poderá identificar se a produção de um aluno se aproxima ou se distancia do que os professores esperam para uma determinada série. Desta forma, sobretudo em casos de distanciamento entre a produção observada e as produções esperadas, a IA poderá se configurar como uma ferramenta pedagógica auxiliar para o professor, permitindo que ele localize os aspectos que necessitam ser focalizados para que um aluno alcance o desempenho esperado para o seu grau de instrução.

Boa parte das nossas atividades no segundo semestre de 2020 se voltou à construção do modelo de execução da IA. A seção seguinte relata o processo de construção do modelo de execução.

2.2 - Primeiros passos para a elaboração da IA

O primeiro passo para a elaboração da IA era conseguir um conjunto inicial de dados que nos permitisse verificar padrões nos "erros" de crianças do Ensino Fundamental e associá-los ao nível de instrução formal das crianças. Recorri ao prof. dr. Lourenço Chacon (UNESP/Marília), que se dedica à aquisição da ortografia há muitos anos e que coordena o "Grupo de Pesquisa Estudos sobre a Linguagem" (GPEL). Esse grupo tem um banco de dados de textos produzidos por crianças do Ensino Fundamental e o prof. Lourenço prontamente se dispôs a nos dar acesso ao banco de dados.

Os textos que constituem o banco - elaborados especificamente para esse fim - foram produzidos por crianças do primeiro ao quinto ano de uma escola pública da cidade de Marília (SP). Resultam de uma tarefa de recontar uma narrativa lida pela professora da turma, em sala de aula. O objetivo inicial dos colegas da UNESP era colher quatro textos de cada criança de cada série, mas isso não foi possível para todas as crianças, porque algumas delas faltaram à aula no dia de uma ou mais de uma coleta. Além disso, foram incluídos no banco apenas os textos das crianças cujos pais assinaram o Termo de Consentimento Livre e Esclarecido. Logo, o número de textos/série não corresponde, necessariamente, ao número total dos alunos daquela série. Quando o prof. Lourenço nos permitiu acesso ao banco, já havia textos em meio digital, que alunos participantes do GPEL tinham digitado, mantendo a ortografia original e, em seguida, redigindo os textos de modo a "traduzir" as hipóteses norteadoras dos "erros" verificados. Entretanto, havia também textos que tinham sido escaneados, mas ainda não estavam em meio digital. Tentamos recorrer a um leitor OCR, que convertesse textos em escrita cursiva para meio digital, mas não foi possível usar um, seja porque nem sempre a letra das crianças era visível ou inteligível, seja porque a distância de

uma produção relativamente à ortografia padrão não permitia que o leitor reconhecesse a palavra. Por isso, eu digitei os textos, seguindo os mesmos critérios adotados pelo pessoal do GPEL, até mesmo como forma de oferecer alguma contrapartida pela cessão do banco.

A interlocução com o prof. Lourenço trouxe outras contribuições para o projeto: por aconselhamento dele, decidimos que seriam tomados para a fase de aprendizado e as primeiras clusterizações textos de terceiro e quinto anos. Tal decisão se justifica pelo fato de essas serem as séries que encerram, respectivamente o primeiro e o segundo ciclos do Ensino Fundamental I e, portanto, envolverem a avaliação das produções das crianças.

Resulta dessa decisão que o *corpus* de treinamento da nossa IA é constituído de 65 textos de alunos de terceiro ano e 63 textos de alunos de quinto ano, perfazendo um total de 128 textos. Aparentemente o *corpus* para treinamento da IA é pequeno, porém é necessário esclarecer que não estamos tomando o texto como unidade de análise, e sim as palavras, o que resulta num conjunto de dados de tamanho razoável para o treinamento inicial.

Como o foco do projeto é a ortografia - e não, ainda, a geração de textos - tomar a palavra como unidade de análise deve nos permitir chegar a outras duas unidades linguísticas importantes da ortografia: a sílaba e o segmento. Mais adiante eu apresento um sistema de etiquetamento que eu elaborei para dar conta do aprendizado de padrões silábicos e fonotáticos pela máquina.

2.3 - Os padrões de "erros" nos textos

Definido o corpus de aprendizagem da IA, o passo seguinte foi analisar cada texto em busca de padrões que pudessem caracterizar os "erros" e, ao mesmo tempo, que pudessem fornecer subsídios para a programação da nossa IA, objetivando inclusive estabelecer padrões associados ao terceiro ou ao quinto anos do Ensino Fundamental.

Antes de passarmos à análise dos "erros", um esclarecimento: daqui em diante eu me referirei às crianças do Ensino Fundamental, em cujas produções baseamos nossa IA, como "escreventes", ou seja, aqueles que escrevem, seguindo a designação proposta por Chacon e Pezarini (2018). Esse nome norteia a escolha do nome da nossa IA: "scriba" é o termo em latim para "escrevente". Dessa forma, também buscamos sinalizar que nossa IA Scriba deve aprender a escrever como crianças do Ensino Fundamental. Passemos, agora, às análises dos textos de cada série:

No terceiro ano é mais difícil apontar textos mais próximos do padrão, diferentemente do 5o. ano, como era de se esperar. Ao mesmo tempo, muitos "erros" que aparecem no quinto ano aparecem também no 3o., embora eu fique com a impressão de que são mais recorrentes no 3o. ano. É o caso do emprego de **um <r> ou um <s> só**, em substituição a <rr> e <ss>, como em "erada", "moreu", "nosa", "buro", "feramenta". É também o caso da harmonia vocálica, como em

"descubriu" ou "asupru" (assoprou). Observação análoga vale para a grafia de /w/ em final de palavra. No 3o. ano vemos produções como "tiu". Alguns sujeitos fazem **hipercorreção,** especialmente **em marca de pretérito perfeito na 3a. pessoa do singular**. Como resultado, encontramos as formas: "acreditol"; "levol"; "pulol"; "mechel"; "perguntol"; "respodel".

Alguns aspectos, por outro lado, parecem característicos de textos dessa série:

1) nasalidade da vogal: não se trata, como no 5o ano, de marcar a nasalidade com <n> ou <m>, quando a norma padrão estabelece justamente o contrário. Trata-se simplesmente de **não marcar a nasalidade da vogal**, como na última sílaba de "tanbei" (também) ou na primeira de "etão" (então) e "leba" (lembrar). Há inclusive casos de escreventes que quase não marcam nasalidade vocálica;

2) <am> x <ão>: esse é um desdobramento do primeiro ponto, mas no 3o. ano vemos que as crianças se confundem quanto à forma ortográfica de 3a. pessoa do plural do pretérito do indicativo ou do futuro do indicativo. Como resultado, temos "resolverão"; "inveterão"; "descobrirão", que são empregadas no passado. Temos "achão" por "acham" e "cãopainha" para "campainha";

3) troca de consoantes desvozeadas (sem vibração de pregas vocais) por vozeadas (com vibração de pregas vocais): "vazendo" (fazendo); "vazer" (fazer) "borta" (porta); "asvavas" (às favas); "so bo gaza" (só por causa); "dipo" (tipo); "vicou" (ficou); "drocando" (trocando). Cabe notar que os três últimos exemplos foram produzidos todos pelo mesmo escrevente;

4) troca de consoantes vozeadas (com vibração de pregas vocais) por desvozeadas (sem vibração de pregas vocais): este caso é o espelho dos exemplos em (3), acima, e compreende produções como "parti" (bati); "enpora" (embora); "secredo" (segredo); "soava" (zoava);"vicitas" (visitas);

5) problemas com dígrafos: ou as crianças **não grafam** os dígrafos - como em "biginhos" (bichinhos); "cupa" (chupa); "escoleu" (escolheu)**;** "brancinha" (branquinha) -, **ou trocam <nh> e <lh>**, como em "armadinha" (armadilha); "porquilho" (porquinho), "canpailha" (campainha); "milha" (minha);

6) grupos consonantais com <r> ou <l>: ou as crianças **desdobram o grupo**, promovendo uma reorganização silábica com a criação de uma nova sílaba - como em "pereparou" (preparou); "eisi polodiu" (explodiu); "ei torou" (entrou); "asucara" (açúcar) - ou as crianças **apagam o <r>** - como em "osoto" (os outros); "povo" (provou"); "cadado" (quadrado); "pecupada" (preocupada); "enveconhados" (envergonhados); "leba" (lembrar); "a sopou" (assoprou); "macelo" (Marcelo); "matelo" (martelo); "atapalhar" (atrapalhar); "pemdeu" (prendeu). Os problemas parecem mais recorrentes com grupos com <r> e esses grupos, na sua maioria, têm <r> apagado.

É preciso comentar, adicionalmente, que em vários dos exemplos do banco, há "problemas" conjugados: em "baguselo", a troca de <r> por <l> se conjuga com a redução do ditongo e a falta de marca de nasalidade da vogal na 3a. sílaba (estou contando do final para o começo da palavra). **A conjugação de "erros" parece muito menos frequente nos textos das crianças do 5o. ano e, por isso, talvez também possa caracterizar os textos do 3o. ano**.

Finalmente, em alguns textos de escreventes de 3o. ano ainda é possível encontrar exemplos de reunião aleatória de letras, um fato esperado sobretudo para textos de escreventes de 1o. ano.

Observando, na sequência, as produções dos escreventes de 5o. ano, e considerando todas as quatro coletas, há textos cuja ortografia está bem mais próxima da norma. Isto não significa que esses textos não tenham tido nenhum caso de desvio dela. Comparativamente com os demais, tiveram muitos menos. Para fazer esse levantamento dos textos mais próximos da norma ortográfica, eu considerei que "erros", como "chícara", "mechedor", "caxinbo", repetidos várias vezes num mesmo texto eram uma única instância do mesmo "erro". Os textos que seguem esse padrão, e que foram produzidos por quatro crianças, são tomados como ponto de partida para caracterizar um escrevente de quinto ano. Assim, os "erros" mais recorrentes nas produções das crianças dessa série são:

1) final de palavra com ditongo terminado em /w/[5], como em "Brasil", "Rio", "partiu": há uma grande flutuação sobre a grafia desse som, mas me parece possível estabelecer uma generalização segundo a qual verbos no passado (3a. pessoa do singular), muitas vezes têm a marca morfológica <-iu> grafada como <-io>. Disso resultam exemplos como "saio" (saiu); "caio" (caiu). Em substantivos, há o fato inverso, então "tio" é grafado recorrentemente "tiu". Por outro lado, não é incomum haver hipercorreção. No quinto ano, ela recai sobre nomes (e não sobre formas verbais). Assim, temos "bacalhal" para "bacalhau";

2) <ss> e <rr> em meio de palavra: comum escreventes usarem um <s> ou um <r> apenas, como em "travesura, "cachoro", "espiro", "terivel". Esse fato, identificado por Chacon e Pezarini (2018) é interessante, porque viola uma norma regrável: sons como [s] e [r], no meio de palavra, entre vogais, são grafados <ss> e <rr>;

3) redução do ditongo, que se verifica tanto no meio como no final de palavra e transpõe para a escrita o que fazemos na fala: "estorou"; "poera"; "tercero"; "bobera"; "pego";

5 As barras sinalizam que a unidade em seu interior é abstrata, i.e., a representação de um som ou um conjunto de sons que compartilham algumas características articulatórias; em suma, os fonemas. Neste caso, /w/ compreende o último som de palavras como "anel" e "chapéu", que exibe tendência a ser produzido exatamente da mesma forma, embora ainda haja falantes de português brasileiro, em algumas poucas regiões, que produzem o último som de "anel" de modo distinto daquele como produzem o último som de "chapéu".

4) harmonia vocálica (influência da qualidade da vogal tônica sobre a pretônica): "puera"; "ispludiu" (este "erro" me parece particularmente interessante, porque a vogal tônica exerce influência sobre as duas pretônicas); "sinti"; "rechunchuda";

5) supressão do <r> que marca infinitivo verbal: transpõe para a escrita prática corrente na fala e é fato quase categórico na escrita das crianças do 5o. ano;

6) "mais" como conjunção adversativa no lugar de "mas": eu arrisco dizer que essa ditongação está quase lexicalizada dessa forma para os falantes de português brasileiro em geral e que essa produção já reflete o fenômeno.

Existem alguns **fatos flutuantes:**

1) marca de nasalidade da vogal: algumas crianças generalizam a marca como <m>, outras como <n>, outras usam ambas. Então, é possível encontrar "vomtade"; "emtão"; "em tam", assim como é possível encontrar "bonba", "caxinbo", "enbora";

2) em alguns casos, marca de gerúndio se reduz a <-no>: "falano", que pode refletir, nos textos, uma característica dialetal dos escreventes;

3) trocas de consoantes oclusivas vozeadas (/b, d, g/) pelas contrapartes desvozeadas (/p,t,k/) transpostas para a escrita: "caximpo". Isto acontece em alguns poucos textos, parecem marcas individuais.

Para além dessas, as trocas de <ch> por <x> (ou o inverso), é bastante frequente: temos "chícara", "mechedor/mecher"; "buxexa", por exemplo.

Cabe acrescentar que alguns dos "erros" produzidos pelos escreventes são sensíveis à estrutura acentual das palavras, como é o caso daqueles motivados por harmonia vocálica (e.g. "ispludiu", "sinti"). Outros são sensíveis à estrutura silábica, como no caso dos grupos consonantais com <r> ou <l> e nos quais verificamos o desdobramento do grupo (e.g. "pereparou", "eisi polodiu"), ou o apagamento da segunda consoante do grupo (e.g. "lemba", "osoto"). Há, ainda, "erros" sensíveis a domínios prosódicos maiores do que a sílaba, como as palavras fonológicas. É o caso de exemplos como "osoto" (os outros) ou "a sopou" (assoprou). Outros erros, finalmente, parecem sensíveis a informação morfológia: assim, a flutuação da grafia de ditongos do tipo /Vw/ - onde V, por hipótese, é qualquer uma das sete vogais orais do português - parece se dar, nos textos de 5o. ano, de modo que verbos tenham o ditongo grafado <io>, enquanto nomes têm grafia alternada entre <Vu> (como em "tio") e entre <Vl> (como em "bacalhal").

A sensibilidade dos "erros" à estrutura silábica e acentual das palavras levou à elaboração de um conjunto de etiquetas para tratamento dos dados. A seção seguinte apresenta e explica as etiquetas.

2.4 - Etiquetas para a estrutura interna das palavras-alvo

O quadro 1, abaixo, reúne as variáveis e os valores atribuídos a cada variável. Neste caso específico, os valores das variáveis são a representação ortográfica para os diversos sons do português brasileiro, para as posições possíveis dos sons no interior das sílabas e para a estrutura acentual das sílabas. As etiquetas são propostas para cada variável elencada.

Quadro 1 - Etiquetas para variáveis

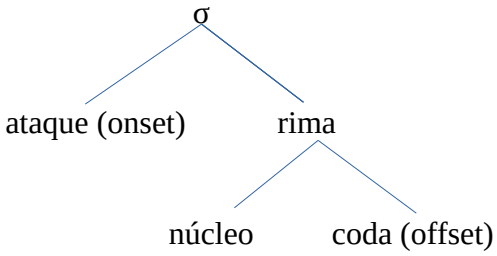
Variável	**Representação ortográfica**	**Etiqueta**
Consoantes oclusivas	p, b, t, d, c, qu, g, gu	O
Consoantes fricativas	f, v, s, ss, c, x, z, ch, j, g	F
Consoantes nasais	m, n, nh	N
Consoantes líquidas	l, lh, r, rr	L
Vogais	i, e, a, o, u, ẽ, ã, õ	V
Ataque simples	O, F, N ou L	SA
Núcleo simples	V	SN
Coda simples	p, t, d, c, g, f, s, z, m, n, l, r	SC
Primeira posição de ataque ramificado	p, b, t, d, c, g, f, v	CA1
Segunda posição de ataque ramificado	l, r, s, m, n	CA2
Primeira posição de núcleo complexo	i, e, a, o, u, ã, õ	CN1
Segunda posição de núcleo complexo	i, u, e, o	CN2
Primeira posição de coda complexa	n, r	CC1
Segunda posição de coda complexa	s	CC2
Sílaba tônica		3
Sílabas pretônica e postônica		1
Sílaba postônica átona final		0

Fonte: A autora.

A razão para o foco sobre a sílaba está no fato de que ela é a unidade linguística em que se verifica o relacionamento entre sons da fala e nos permite fazer previsões acerca do comportamento desses sons. Assim, por exemplo, permite prever por que, em português, temos uma sílaba como "por", em "porta", e não uma sílaba "rt". A importância dessa previsão para aprendizado de máquina é óbvia: como falantes de português, nós conhecemos a estrutura interna das sílabas da nossa língua e, portanto, não temos problema em estabelecer as fronteiras silábicas. Mas uma máquina não conhece a estrutura interna das sílabas do português. Se, como observado na seção 2.3, a sílaba é o domínio linguístico em que se verificam vários "erros", é preciso ensinar à nossa IA como é a estrutura interna das sílabas do português. Dessa forma, a Scriba terá subsídios para aprender os

desvios, a partir das formas canônicas e, em consquência, reproduzi-los ou caracterizar produções em função do nível de escolaridade do escrevente.

Para uma melhor compreensão das etiquetas propostas, são necessários alguns esclarecimentos. O primeiro diz respeito à estrutura interna da sílaba, que a Linguística concebe da seguinte forma:



Nessa estrutura, o constituinte obrigatório é o núcleo que, no português, é ocupado apenas por vogal. Os constituintes “ataque/onset” e “coda/offset” são opcionais e preenchidos por consoantes. Nas etiquetas, esses constituintes são sinalizados, respectivamente como SN, SA; SO, onde "S" corresponde a "sílaba" e a letra seguinte, a um dos três constituintes mencionados.

Além disso, nos três constituintes – ataque, núcleo e coda – os termos “ramificado” ou “complexo” remetem a um mesmo fenômeno, que é a presença de mais de uma unidade sonora no constituinte silábico. Assim, um ataque complexo remete a um início de sílaba com duas consoantes (como em “**pr**aga”); mesma observação vale para uma coda complexa (como em “**pers**pectiva”). No caso do núcleo complexo (ou ramificado), temos a presença de duas vogais (como em “c**ai**xa”, “p**ou**co[6]”). Nas minhas etiquetas, eu uso apenas o termo “complexo”, para simplificar as etiquetas. E coloco o adjetivo precedendo os substantivos “ataque”, “núcleo” e “coda” (mesmo padrão exposto em (4)). A posição da unidade sonora no grupo consonantal ou vocálico é marcada pelos índices 1, 2. Sobre os **ataques complexos,** é preciso observar que a segunda unidade é geralmente uma líquida (l, r). Ainda assim, o português não tem encontro <dl> e o encontro <tl> se restringe a empréstimos do grego (“atlas”, “atleta” e derivados). Existem palavras onde a segunda unidade é a consoante fricativa <s>, como em “psicólogo” (e seus derivados) ou empréstimos como “tsuru”, “tsunami”. Há alguns poucos casos em que a segunda unidade é a consoante nasal <n>, como em “pneu” (e derivados) ou “pneumonia”. Embora na fala produzamos uma vogal entre essas duas consoantes e formemos, assim, uma sílaba adicional nessas últimas palavras (caso de “peneu”), eu estou considerando a forma ortográfica, que é nosso foco aqui. Quanto aos núcleos complexos, característicos de ditongos, cabe comentar que o segundo membro é sempre <i> ou <u>, como em “c**ai**s” ou “n**au**”. Se existem duas consoantes iguais contíguas (como em “veem”, cada vogal

6 Essa análise é uma dentre as possíveis, porque há autores que consideram o segundo membro dos ditongos uma consoante, e não uma vogal. Adotei essa análise apenas para manter uma uniformidade epistemológica com outros estudos nos quais a Scriba se baseia.

constitui, sozinha, o núcleo de uma sílaba, o que resulta em 2 sílabas e não apenas 1). Mesma observação vale para o caso de a segunda vogal ser diferente da primeira, mas não ser nem <i>, nem <u>, como em "**ao**risto" ou "ide**ia**". Finalmente, no caso das codas complexas, temos que o seu segundo membro se restringe à consoante <s>, como em "**trans**porte" ou "**pers**pectiva".

As etiquetas contemplam também **restrições fonotáticas** do português brasileiro. As restrições fonotáticas dizem respeito às sequências de sons permitidas ou não numa língua. Assim, e.g., uma sequência como <shr> não é permitida em nossa língua, embora seja noutras, como o inglês. Para propor as etiquetas contemplando as restrições fonotáticas, considerei que o inventário de consoantes licenciadas em ataque é maior do que o inventário de consoantes em coda. Além disso, previ que a sequência de consoantes em ataque e coda complexos não é aleatória, mas obedece a uma "escala de sonoridade". Ohala e Kawasaki-Fukumori (1997), e.g., definem a escala de sonoridade da seguinte maneira: oclusiva < fricativa < nasal < líquida < vogal. A escala, grosso modo, se constroi sobre o parâmetro "grau de obstrução à passagem do fluxo de ar no trato vocal", que é correlacionado à saliência perceptual das consoantes, ou à sua sonoridade. Oclusivas são produzidas necessariamente através de bloqueio total à passagem do fluxo de ar no trato vocal. Por isso, são menos salientes perceptualmente e, portanto, estão no último lugar do *ranking* de sonoridade. Vogais são produzidas sem oferecer bloqueio à passagem do ar no trato vocal, o que faz delas sons mais salientes em termos perceptuais. Por isso ocupam o topo da escala. Contemplar essas restrições fonotáticas é especialmente necessário para um algoritmo que faça a separação silábica das palavras do português, pois se pode prever as estruturas silábicas do português a partir das restrições fonotáticas mais gerais em (10) e das mais específicas nos pontos anteriores.

As etiquetas contemplam, ainda, o **acento,** que necessita de uma abordagem específica, pois tem natureza diferente dos sons, espalhando-se por uma sílaba inteira. Por essa razão, e considerando que temos níveis de acento no português - mais ou menos intensos - elaborei as etiquetas recorrendo à proposta de Camara Jr. (1970) que, embora seja bem antiga, continua aceita na literatura e serve bem aos nossos propósitos. Como se vê no Quadro 1, preveem-se os índices 0, 1 e 3. O índice "2" falta no quadro de propósito: ele é reservado para o que a literatura chama "acento secundário". Ele é uma espécie de "vestígio" do acento principal (tônico) em palavras derivadas: em "cafezinho", a tônica é <zi>, mas <fe> tem intensidade intermediária à tônica e às demais sílabas. Fato parecido acontece em "economicamente" com a sílaba <no> e nos demais derivados com -mente. Pensei nessa estratégia prevendo a possibilidade de, em algum momento, a gente lidar com esse tipo de fato. Uma observação adicional, para melhor explicação sobre os índices e as variáveis: a sílaba tônica é a mais intensa, por isso recebe o maior grau de acento (3). A

sílaba pretônica e a postônica, que precedem e sucedem a tônica, recebem índice igual, mas mais baixo do que a tônica. Numa palavra como “abóbora”, a sílaba <bó> receberia nível 3 de acento, por essa proposta; as sílabas <a> e <bo>, nível 1. A sílaba <ra>, que segue a tônica e ocorre no final de palavra, recebe nível zero de acento. A sílaba átona final, aliás, por ser prosodicamente mais fraca, pode ser apagada, i.e., pode não ser produzida na cadeia da fala. No caso da Scriba, essa informação pode ser relevante, considerando que os escreventes - sobretudo nas séries iniciais - podem levar para a escrita características que percebem na cadeia da fala.

Um último comentário sobre as etiquetas dispostas no Quadro 1: considerando a observação de Chacon e Pezarini (2018), segundo a qual as **classes de sons** desempenham papel relevante nos “erros”, criei as etiquetas O, F, N, L e V. Essas etiquetas podem ser relevantes caso queiramos contemplar a variável "classes de sons" no nosso sistema, de modo que ele possa dar conta de que uma oclusiva é trocada por outra oclusiva, por exemplo nos casos em que os escreventes "trocam" <f> por <v> em "avavas" (por "às favas", na expressão "vá às favas"), ou "trocam" <b> por <p> em "enpola" (por "embora").

Finalmente, cabe mencionar dois aspectos não contemplados nas etiquetas:

1) não se consideram o número de sílabas das palavras, porque Chacon e Pezarini (2018) não observaram influência desse fato sobre os "erros" dos escreventes. Isso não significa, porém, que a informação não possa ser acrescentada, caso seja necessária em algum momento;

2) em comunicação pessoal, Chacon comentou sobre a importância de as etiquetas considerarem, em algum momento, as codas fonológicas e as morfofonológicas, isto é, as codas que só carregam informação sonora e aquelas que, além da informação sonora, carregam também informações morfológicas, como classe gramatical ou tempo verbal. Eu não sei como automatizar essas etiquetas, não da forma como eu previ as etiquetas expostas nesta seção. É que a diferenciação entre os tipos de coda me parece requerer a construção de um *parser*, que possibilite ao sistema saber se uma palavra é verbo ou substantivo, e.g., de modo que ele consiga distinguir que <r> em “isopor” é uma coda fonológica, mas <r> em “compor” é coda morfofonológica, porque o som que a letra representa também é marca de infinitivo verbal. Observação análoga vale para <m> em formas verbais (e.g. 3ª pessoa plural pretérito perfeito). Não é impossível elaborar o *parser*, claro, mas decidimos não enveredar por essa tarefa neste momento, porque avaliamos que seria possível construir a Scriba sem ele e porque, elaborá-lo nos faria desviar do caminho a ser focalizado.

As etiquetas propostas nos permitem tratar os dados como no Quadro 2, a seguir:

Quadro 2 - Exemplos de palavras etiquetadas

ETIQUETAMENTO

3º Ano

Exemplares	Etiquetas com tipos de consoantes e níveis de acento
1 sílaba	
mau	[(SAN)(CN1)(CN2)]3
sai	[(SAF)(CN1)(CN2)]3
2 sílabas – ox	
senhor	[(SAF)(SN)]1[(SAN)(SN)(SCL)]3
inflei	[(SN)(SCN)]1[(CA1F)(CA2L)(CN1)(CN2)]3
2 sílabas – par	
porco	[(SAO)(SN)(SCL)]3[(SAO)(SN)]0
crânio	[(CA1O)(CA2L)(SN)]3[(SAN)(CN1)(CN2)]0
3 sílabas - ox	
arrombar	[(SN)]1[(SAL)(SN)(SCN)]1[(SAO)(SN)(SCL)]3
derrubei	[(SAO)(SN)]1[(SAL)(SN)]1[(SAO)(CN1)(CN2)]3
3 sílabas- par	
bochecha	[(SAO)(SN)]1[(SAF)(SN)]3[(SAF)(SN)]0
açúcar	[(SN)]1[(SAF)(SN)]3[(SAO)(SN)(SCL)]0
3 sílabas-prop	
xícara	[(SAF)(SN)]3[(SAO)(SN)]1[(SAL)(SN)]0
vítima	[(SAF)(SN)]3[(SAO)(SN)]1[(SAN)(SN)]0
4+ sílabas par	
vovozinha	[(SAF)(SN)]1[(SAF)(SN)]2[(SAF)(SN)]3[(SAN)(SN)]0
aniversário	[(SN)]1[(SAN)(SN)]1[(SAF)(SV)(SCL)]1[(SAF)(SN)]3[(SOL)(CN1)(CN2)]0

5o Ano

3 sílabas - ox	
refeição	[(SAL)(SN)]1[(SAF)(CN1)(CN2)]1[(SAF)(CN1)(CN2)]3
construiu	[(SAO)(SN)(CC1N)(CC2F)]1[(CA1O)(CA2L)(SN)]1[(CN1)(CN2)]3
3 sílabas- par	
poeira	[(SAO)(SN)]1[(CN1)(CN2)]3[(SAL)(SN)]0
presunto	[(CA1O)(CA2L)(SN)]1[(SAF)(SN)(SCN)]3[(SAO)(SN)]0
4 sílabas - par	
resfriado	[(SAL)(SN)(SCF)]1[(CA1F)(CA2L)(SN)]1[(SN)]3[(SAO)(SN)]0
descobriram	[(SAO)(SN)]1[(SCF)(SAO)(SN)]1[(CA1O)(CA2L)(SN)]3[(SAL)(SN)(SCN)]0

Fonte: A autora.

O quadro 2 dispõe palavras tratadas com as etiquetas expostas no Quadro 1. Todas essas palavras foram retiradas de produções dos escreventes relativas à primeira coleta, consistia na

reescrita de "A Verdadeira História dos Três Porquinhos". Na coluna com os "exemplares", as palavras encontradas em textos de escreventes de 3o. ano estão presentes também nos textos dos escreventes do 5o. ano, mas o inverso não é verdadeiro, o que sugere que número de sílabas das palavras e complexidade interna das sílabas podem igualmente ser associados ao nível de instrução dos escreventes.

Um esclarecimento sobre a leitura das etiquetas: tomemos a palavra "vovozinha" como exemplo. A ela foi atribuída a etiqueta em (i):

(i) [(SAF)(SN)]1[(SAF)(SN)]2[(SAF)(SN)]3[(SAN)(SN)]0.

Essa notação traz, entre parênteses, informações sobre um som específico. Assim, (SAF) corresponde à consoante fricativa (F) do ataque silábico (SA); (SN) corresponde ao núcleo da sílaba. As fronteiras de cada sílaba são marcadas por colchetes e, logo após as sílabas está informado o nível de acento. para a sílaba [(SAF)(SN)] o nível de acento é 1, já que se trata de uma pretônica. Na sequência, temos nova sílaba, cuja consoante inicial é a mesma da sílaba precedente. Essa sílaba também se encerra com o núcleo silábico, mas recebe acento de nível 2. Isto porque, como mencionado mais acima, nesta seção, temos aí o acento secundário, que é resquício do acento principal da palavra "vovó", base para a derivação "vovozinha". Em seguida, temos nova fricativa em ataque silábico, seguida por uma vogal que ocupa o núcleo da sílaba seguinte. Essa sílaba, como vemos, recebe nível 3 de acento, porque carrega o acento primário da palavra. Finalmente, temos uma sílaba cujo primeiro som corresponde a uma consoante nasal (N) que ocorre no ataque (SA) e é seguida por um núcleo silábico (SN). Como essa sílaba é átona e encerra a palavra, sua intensidade é muito tênue, o que leva à atribuição do nível 0 de acento.

Um problema dessas etiquetas é que elas não especificam qual vogal ocorre em núcleo silábico e essa informação pode ser importante para que a Scriba aprenda exemplos de "erros" que envolvem harmonia vocálica, como "rechunchuda" (por "rechonchuda") ou "ispludiu" (por "explodiu"). Por isso pensei em etiquetas adicionais: "fr" para vogais frontais, como todas as vogais da palavra "hélice" ou as vogais das sílabas átonas de "elefante"; "ct" para vogais centrais, como todas as vogais de "batata"; "pt" para vogais posteriores, como todas as vogais de "urubu", todas as vogais de "vovó" e as vogais de "povo". Há, ainda, um detalhe adicional na articulação das vogais: elas variam em função da altura da mandíbula, de modo que podem ser articuladas com a mandíbula mais alta - e a boca mais fechada, como a vogal <i> em "ícone" - ou mais baixa - e a boca mais aberta, como acontece com as vogais de "batata". Existem graus intermediários a esses, no que concerne à abertura, e que resultam em vogais meio-fechadas - como <e> em "mês" e como <o> em "vovô - e em vogais meio-abertas - como <é> em "café" e como <ó> em "cipó". Então, a proposta é estabelecer graus de abertura da mandíbula, desde 1, mais fechado, até 3, que

caracterizaria as vogais meio-abertas. Esses graus se combinariam com as etiquetas "fr", "ct", "pt". Seria possível - mas aparentemente desnecessário - conceber um grau 4 para dar conta da vogal <a>. Mas como ela é a única vogal central no português brasileiro, especificar seu grau de abertura parece redundante. Feitas essas considerações, as etiquetas da palavra "vovozinha" podem ser reescritas como em (ii):

(ii) [(SAF)(SNpt2)]1[(SAF)(SNpt3)]2[(SAF)(SNfr1)]3[(SAN)(SNct)]0

Estabelecido o *corpus* para o aprendizado da Scriba e oferecido um tratamento inicial aos dados que constituem o banco, pudemos passar à programação da IA.

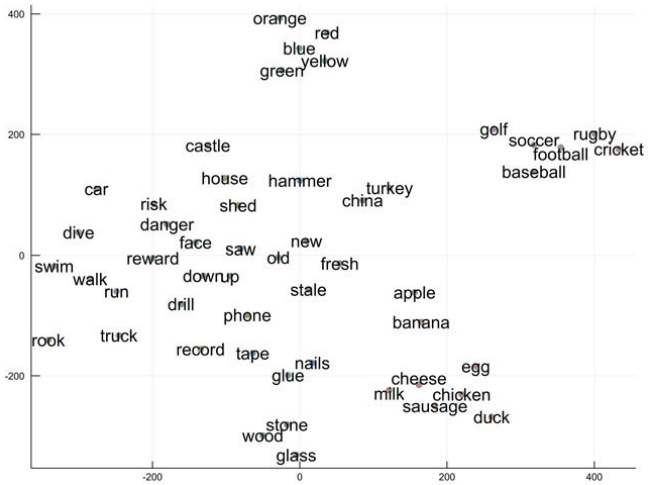
3 - Arquitetura da IA

Primeiramente, é necessário comentar que a Scriba está sendo construída pelo mestrando Wanderlan Carvalho, orientado pelo prof. dr. Fabiano Silva e co-orientado por mim. Dados os devidos créditos, nesta seção eu faço uma breve apresentação inicial sobre as técnicas disponíveis para processamento de linguagem natural e construção de um modelo de linguagem, quando se considera o aprendizado de máquina para, então, comentar os direcionamentos para nossa IA, nosso "modelo de escrevente".

Uma ressalva: em razão da redefinição do meu projeto, por um lado, e das dificuldades impostas pela pandemia do novo coronavírus, por outro, a Scriba continua em desenvolvimento. Como eu comento na seção introdutória deste relatório, o objetivo é continuar a interlocução iniciada com meu estágio pós-doutoral e não só - obviamente - concluir a Scriba mas trabalhar em possíveis desdobramentos dela.

Voltando, então, às técnicas para processamento de linguagem natural, White *et al.* (2019). notam que as *word embeddings* - também chamadas *word vectors* - são representações vetoriais de uma palavra e consituem uma inovação que possibilitou ao aprendizado de máquina assumir a vanguarda do processamento de linguagem natural. Com elas criam-se vetores para uma palavra a partir de características salientes dessa palavra, como, e.g., o sentido.

Figura 1 - Exemplo de word embedding criada a partir do Fast-Text[7]



Fonte: White et al. (2019:38)

A Figura 1 ilustra um exemplo de um *word embedding*, no qual os grupos de palavras são constituídos em função de semelhança de sentidos. Colocando de outro modo, as palavras de um mesmo grupo pertencem a um mesmo campo semântico. Assim, por exemplo, na porção superior do gráfico é possível reconhecer um grupo de palavras pertencentes ao campo semântico "cores". Por outro lado, à direita, no canto inferior do gráfico podemos reconhecer um campo semântico "alimentos", que une palavras como "cheese", "sausauge", "egg".

A elaboração de vetores de palavras (*word vectors*) é um passo considerável para o processamento de linguagem natural e para a tarefa de modelamento da linguagem porque eles permitem prever, e.g., numa sentença, qual a palavra seguinte, consideradas as precedentes. Assim, numa sentença como "Hoje, no almoço, eu vou comer _____" há uma probabilidade alta de que a última palavra seja "restaurante". Mas ela não seria um verbo qualquer, por exemplo.

Um tratamento usual para o modelamento da linguagem é um tratamento estatístico baseado em trigramas, que envolve o cômputo de trios de palavras num *corpus*. A partir da probabilidade de o trio acontecer em conjunto, pode-se condicionar as duas primeiras palavras para obter uma probabilidade condicional da terceira. Este procedimento atende à premissa de Markov, segundo a qual o estado seguinte de um sistema depende apenas de seu estado atual e de que o estado pode ser descrito pelas duas palavras anteriores. De modo geral, trata-se, portanto, de definir o estado de Markov para que ele contenha qualquer valor de "n" desejado. Cabe, porém, observar que essa premissa é uma aproximação e que essa aproximação pode perder a informação-chave porque, na sentença tomada como exemplo, podemos ter algo como "Hoje, no almoço, eu vou comer livros" e

7 A título de esclarecimento - e mesmo porque eu não vou abordá-lo neste relatório, o Fast-Text é uma técnica de word-embedding desenvolvida pela equipe de pesquisadores em IA do Facebook.

embora a sentença seja sintaticamente bem formada, ela não faz sentido. Ou seja, a premissa de Markov torna tratável a linguagem e permite antever algum modelamento para ela, mas o modelo tem de "aprender" que numa sentença como a do exemplo, o campo semântico engengrado pelo verbo "comer" remeta a alimentos. Considerando, por outro lado, que há outras palavras que poderiam preencher a lacuna, e que não estão propriamente incluídas no campo semântico "alimentos", como, e.g., "pouco", é necessário um método que modele a linguagem contemplando mais informação linguística e, portanto, seja mais flexível. Uma rede neural consegue "aprender" esses atributos e, por isso, pode ser usada para elaborar um modelamento de linguagem.

Assim, e.g., Bengio *et al. (apud* WHITE et al, 2019) propõem uma rede neural baseada não em tri-gramas - o que seria possível -, mas em n-gramas, o que a torna mais avançada. Nela, os n-gramas são hiperparâmetros do modelo, que também é dotado de uma camada *bypass.* Tal camada permite que o *input* afete diretamente o *output* da rede, sem a mediação de camadas escondidas. Outra contribuição importante desse modelo é que, nessa rede, o vocabulário de entrada e o de saída não precisam ser os mesmos. Sendo o vocabulário de saída um subconjunto do vocabulário da entrada, o tempo necessário para o treinamento da rede diminui. O modelo de Bengio *et al.* é o primeiro a usar as redes neurais com vetor de representação de palavras e atua de modo muito parecido com os n-gramas tradicionais, já que, a cada instante temporal, a janela desliza para a frente e a próxima palavra é prevista com base no contexto das palavras anteriores. Os autores mencionam rapidamente, em seu modelo, a possibilidade de uso de redes neurais recorrentes (RNN) em substuituição à rede que eles propõem.

É o que Mikolov *et al.* (2010, *apud* WHITE ET AL. , 2019) fazem: ao utilizarem uma RNN, eles eliminam a premissa de Markov de que uma janela finita de palavras precedentes forma o estado atual da rede. Ao contrário, o estado é aprendido e armazenado pela rede. Uma técnica associada ao funcionamento dessa RNN é o CBOW (*Continuous Bag of Words*), que não é semelhante a um *bag of words*, apesar da denominação, mas que, a exemplo outras representações de palavras - como o skip-gram ou o GloVe - que não considera a ordem das palavras de contexto. O CBOW toma como input uma janela de contexto ao redor de uma palavra central que foi pulada, e tenta prever qual palavra é essa. As janelas podem se localizar em ambos os lados da palavra pulada, sem que seja necessário um procedimento linear.

O "espelho" do CBOW são os modelos do tip *skipgram*, em que, a partir de uma palavra central, preveem-se as palavras do contexto. O *word2vec* é um algoritmo do tipo *skipgram* bastante conhecido. Os modelos *skipgrams* tomam uma palavra isolada como *input e seu output* é uma função *softmax* que verifica a probabilidade de cada palavra do vocabulário ocorrer no contexto da palavra do *input*. Isto pode ser indexado para se obter a probabilidade individual de uma palavra

ocorrer usualmente num modelo de linguagem. O objetivo desse procedimento é maximizar as probabilidades de que os *outputs* observados ocorram, de fato, num dado contexto.

É preciso acrescentar que as técnicas de *word embedding* podem ser usadas satisfatoriamente para expressar analogias utilizando álgebra linear. As analogias podem ser semânticas ou sintáticas, ou seja, o modelo deve conseguir prever que marido está para esposa como bode está para cabra, ou que marido está para maridos, assim como esposa está para esposas. As tarefas de analogia podem ajudar a identificar o tipo de semelhanças capturadas pelas *word embeddings*, mas um estudo de Mikolov *et al.* (2013a, *apud* WHITE *et al.*, 2019) concluiu que, enquanto o CBOW tem desempenho fraco para tarefas de analogias semânticas, mas bom desempenho para tarefas de analogias sintáticas, os *skipgrams* têm bom desempenho em ambas as tarefas, embora não sejam tão bons quanto o CBOW para tarefas sintáticas.

3.1 - *Scriba:* uma IA aplicada ao ensino de ortografia

A *Scriba* está sendo elaborada a partir de uma RNN que emprega representações vetoriais para cada palavra do *corpus*. O objetivo é que ela consiga agrupar "erros" em torno da grafia normatizada de cada item. Para isso, a técnica escolhida para a *word-embedding* foi o word2vec que, como comentado na seção precedente, toma uma forma isolada e oferece, como *input*, formas relacionadas a ela. Permite, adicionalmente, maximizar as probabilidades de ocorrência dos *outputs* observados.

Como as *word-embeddings* permitem expressar analogias entre formas, a hipótese norteadora da elaboração da *Scriba* é de que deverá ser possível traçar analogias entre formas no que concerne também às possíveis formas ortográficas de uma palavra. Além disso, como o word2vec possibilita calcular a distância entre esses pontos do vetor deverá ser possível prever quais formas de "erros" são mais próximas ou mais distantes da forma ortográfica normatizada. Conforme mencionado na seção 2.3, os "erros" de escreventes de 3o. ano parecem se caracterizar pela conjugação de desvios variados, diferentemente dos "erros" de escreventes de 5o que parecem seguir uma tendência a exibir um desvio específico. Portanto, espera-se que haja uma distância maior entre os "erros" e as formas normatizadas em produções de 3o. ano, comparativamente às produções de escreventes do 5o. ano. Confirmada essa hipótese, poderemos dar conta do nosso objetivo de oferecer aos professores uma ferramenta de avaliação das produções de um escrevente: a *Scriba* poderá analisar novos "erros" e apontar se eles se aproximam mais do padrão de "erros" do terceiro ou do quinto ano. Como consequência, em casos desviantes, i.e., em casos nos quais o padrão se aproxime dos erros de 3o. ano, em produções de um escrevente do 5o. ano, o professor poderá pensar ações

individualizadas para levar o escrevente a produzir formas mais próximas daquelas esperadas para seu nível de instrução.

Na fase atual de desenvolvimento da *Scriba*, estão sendo feitos treinamentos visando ao aprendizado da RNN a partir da qual nossa IA se desenvolve. Para o treinamento, Wanderlan tem usado vetores pré-treinados para *word-embedding*. Há vetores pré-treinados disponíveis para CBOW, wod2vec, Fast-Text, por exemplo. A vantagem deles é que eles são treinados a partir de um conjunto de dados muito maior do que aqueles a que as pessoas, no geral, têm acesso. Considerando, e.g., que o Fast-Text é uma técnica desenvolvida pelos especialistas em IA do Facebook, o conjunto de dados nos quais os seus vetores se baseiam é obviamente muito grande. Esses vetores pré-treinados podem, no caso da *Scriba*, ser usados como valores iniciais da RNN.

Wanderlan tem usado vetores pré-treinados disponibilizados pelo Núcleo Interinstitucional de Linguística Computacional (NILC) da USP. O Núcleo tem um repositório de vetores de palavras (geralmente de 300 dimensões) para o português brasileiro e o português europeu, para serem usados em processamento de linguagem natural e aprendizado de máquina. Segundo Hartmann *et al.* (2017), foram treinados 31 modelos de *word-embedding* que usam FastText, GloVe, Wang2Vec e Word2Vec. O Dentre os métodos disponibilizados, Wanderlan tem recorrido aos word-embeddings que utilizam Fast-Text e Word2vec e buscado avaliar os métodos na tarefa de clusterização dos "erros" e das formas normatizadas. É necessário reforçar o comentário de que os métodos são utilizados para analogias sintáticas e semânticas. Por isso, a manipulação das formas ortográficas é um desafio inédito.

Na *Scriba*, a arquitetura da rede deve lidar com cinco camadas que englobam: palavras grafadas conforme o original ("erros"); sequência etiquetada; separação de sílabas (vide item 2.4); estrutura interna das sílabas; níveis de acento das sílabas. As quatro camadas - para além dos erros - foram expostas e comentadas nos Quadros 1 e 2 da seção 2.4. A sugestão do prof. Fabiano é que as camadas possam ser ligadas e desligadas no sistema para cada problema ou contexto com que a RNN for lidar. Dessa forma, as palavras ou sequências delas poderiam ser agrupadas em função de cada um dos eixos.

Em alguma medida, esta estratégia de estrutura da IA adota uma arquitetura semelhante à que sugerem Novais e Paraboni (2013)

> "O sistema toma como *input uma representação abstrada de uma sentença (cf. seção 3.2). Para estabelecer concordância entre os constituintes da sentença, os valores faltantes dos inputs são ajustados e/ou complementados com o auxílio de um banco de dados lexicais*

> *descrito em [9]. O resultado é uma representação totalmente especificada para a sentença, tomada como input do próximo módulo, que sobregera produções alternativas (cf. seção 3.3) baseadas nas restrições de linearização (também obtidas a partir de informação lexical) e, opcionalmente, informação de sinônimos obtidas num dicionário de português brasileiro [10]. Finalmente, o conjunto de candidatos possíveis é submetido a um modelo de linguagem que filtra os dados e seleciona o output mais plausível da sentença."* (NOVAIS, PARABONI, 2013: 137)[8]

Substituindo "sentença" por "palavra" e "concordância entre os constituintes da sentença" por sequências de letras na palavra", temos um procedimento muito parecido com o que descrevo acima. Como consequência dessa arquitetura, a IA aprenderia a produzir os mesmos "erros" que os escreventes de 3o e 5o anos produzem e, para além da avaliação da competência do aluno, que mencionei no início desta seção, a IA pode sinalizar os principais "erros" em cada série e, com isso, podemos atingir nosso outro objetivo, que é o de oferecer diretrizes para o ensino da ortografia.

Mas este é uma etapa posterior da elaboração da IA e que continua para além do encerramento formal do estágio de pós-doutorado. Por ora, é preciso avaliar a técnica de *word-embedding* mais adequada para tratar dos dados com os quais trabalhamos.

4 - Considerações finais

Neste relatório eu busquei apresentar desde a motivação para a criação da *Scriba* até o estado atual da IA, que consiste em avaliar o melhor método de *word-embedding* para a tarefa de aprendizado dos "erros" ortográficos a partir dos quais poderemos construir um modelo de escrevente. Construir modelos computacionais que nos permitam entender como a linguagem humana é processada pode trazer muitas contribuições, de diversas ordens, porque podemos entender as hipóteses subjacentes ao processamento da linguagem em seus vários níveis e, a partir daí, propor soluções para problemas variados.

Construir uma IA capaz de elaborar um modelo de escrevente a partir dos "erros" ortográficos produzidos por crianças de 3o. e 5o. anos do Ensino Fundamental deverá auxiliar professores na

8 No original: "*The system takes as an input an under-specified abstract sentence representation (cf. Sect. 3.2). In order to establish agreement between sentence constituents, missing input values are adjusted and/or complemented with the aid of a lexical database described in [9]. The result is a fully specified sentence representation taken as the input to the next module, which over-generates alternative realisations (cf. Sect. 3.3) based on linearisation constraints (also obtained from lexical information) and, optionally, synonymy information taken from a thesaurus of Brazilian Portuguese [10]. Finally, the set of possible candidates is submitted to a language model filter and the most likely outputsentence is selected.* (NOVAIS, PARABONI, 2013: 137). A tradução é minha.

compreensão desses "erros", e que passa por entender as hipóteses que os norteiam e a identificar padrões característicos de uma ou outra séries. Esta compreensão é particularmente importante porque deve permitir que os professores direcionem práticas pedagógicas para atender às necessidades de alunos cujas produções se caracterizam como produções de outras séries, e não da sua própria. Isso é obviamente relevante nos casos de crianças de 5o. ano que apresentam produções características de crianças de 3o. ano. A classificação das produções pela Scriba pode auxiliar os professores a compreenderem e avaliarem melhor o nível de seus alunos, quanto ao aprendizado da ortografia. Este é um passo importante na alfabetização das crianças, já que a compreensão sobre o funcionamento das relações entre sons da fala e grafema é critério para que se considerem alfabetizados os alunos.

Além da classificação das produções em função do momento de sua instrução formal pelos "erros" produzidos, a compreensão dos "erros" e das hipóteses que os norteiam deve permitir que se proponham diretrizes para o ensino da ortografia, talvez obedecendo à gradiência postulada por Chacon e Pezarini (2018).

Como eu também comentei, a elaboração da Scriba se encontra em fase inicial: já foi oferecido um tratamento inicial a um conjunto de dados, a partir dos quais é possível programar a IA, mas para isso é preciso avaliar qual técnica de *word-embedding* dá conta mais satisfatoriamente dos nossos propósitos. Como decorrência, e voltando a uma afirmação do início deste relatório, a interlocução proporcionada pelo pós-doc com os colegas do DInf e, em especial com o prof. dr. Fabiano Silva continua para além desse meu estágio. E só será possível porque me foi permitido realizar o estágio pós-doutoral na UFPR. O que esperamos é que a elaboração da *Scriba* seja o começo de uma parceria duradoura. Para tanto, estamos prevendo outras atividades, como a criação de interlocução de alunos de Computação e de Linguística, através da criação de um grupo de estudos e atividades de orientação e co-orientação de alunos das duas áreas tanto na Iniciação Científica como na pós-graduação.

5 - Outras atividades

Nesta seção, listo atividades complementares que realizei em paralelo ao projeto de pós-doutorado:

1) participação nas aulas da disciplina de Ciência de Dados, ministrada pelo prof. dr. André Grégio: comecei assistindo às aulas presenciais, interrompidas pela pandemia, e retomei as aulas quando voltaram a ser ofertadas, em janeiro, durante novo período de Ensino Remoto Presencial;

2) participação nas aulas da disciplina de Interação Humano-Computador, ministrada pelo prof. dr. Roberto Pereira, durande período de Ensino Remoto Presencial, no final de 2020 e início de 2021;

3) curso online de programação em Python, durante os meses de abril a julho de 2020;

4) publicação do artigo "42 é a resposta. Qual é a pergunta sobre a relação entre Computação e Linguística e tudo o mais?" na Revista SBC Horizontes, de agosto/2020;

5) co-orientação do projeto de mestrado de Wanderlan Carvalho, que trabalha com processamento de linguagem natural, sob orientação do prof. dr. Fabiano Silva.

Referências

CAGLIARI, L. C. Alfabetização e Linguística. Ed. Scipione, São Paulo, 1989.

CAMARA JR., J.M. *Estrutura da língua portuguesa*. Petrópolis: Vozes, 1970.

CHACON L.; PEZARINI I.O. Gradiência na correspondência fonema/grafema: uma proposta de caracterização do desempenho ortográfico infantil. In: César ABP, Seno MP, Capellini SA. *Tópicos em Transtornos de Aprendizagem: Parte VI*. Ribeirão Preto: Booktoy Livraria e Editora, 2018.

HARTMANN, N.; FONSECA, E.; SHULBY, C.; TREVISO, M.; RODRIGUES, J.; ALUISIO, S. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks, Disponível em arXiv:1708.06025. Acesso em 02/03/2021.

LARKIN, J.H.; CHABAY, R.W. (orgs.) *Computer-assisted instruction and intelligent tutoring systems – shared goals and complementary approaches*. New Jersey: Lawrence Erlbaum Associates, 1992.

LEMLE, M. A tarefa da alfabetização: etapas e problemas no português. Letras Hoje, PUC/RGS, 15, (4), 1982.

MORAIS, A.G.; LEITE, T. M. S. B. R. A escrita alfabética: por que ela é um sistema notacional e não um código? Como as crianças dela se apropriam?. In: *Pacto Nacional pela Alfabetização na Idade Certa: a aprendizagem do sistema de escrita alfabética. ano 1, unidade 3* – Ministério da Educação, Secretaria de Educação Básica, Diretoria de Apoio à Gestão Educacional. Brasília, 2012.

MORAIS, A. G. *Ortografia: ensinar e aprender*. Ed. Ática, São Paulo, 3. ed. 2000.

NOVAIS, E.M.; PARABONI, I. Portuguese text generation using factored language models. *Journal of the Brazilian Computer Society*, 19:135–146, 2013. DOI 10.1007/s13173-012-0095-1

OHALA, J.; KAWASAKI-FUKUMORI, H. (1990). Alternatives to the sonority hierarchy for explaining segmental sequential constraints. In S. Eliasson; E.H. Jahr (orgs.) *Language and its ecology - Essays in memory of Einar Haugen.* Berlin: Mouton de Gruyter: 343-365, 1997.

SILVA, Adelaide Hercília Pescatori. 42 é a resposta. Qual é a pergunta sobre a relação entre Computação e Linguística e tudo o mais? **SBC Horizontes**, agosto 2020. ISSN: 2175-9235. Disponível em <http://horizontes.sbc.org.br/index.php/2020/08/42-e-a-resposta-qual-e-a-pergunta-sobre-a-relacao-entre-computacao-e-linguistica-e-tudo-o-mais/>. Acesso em: 05/03/2021.

WHITE, L.; TOGNERI, R; LIU, W.; BENNAMOUN, M. *Neural representations of natural language*. Singapore: Springer, 2019.