EFFE-ON – *CORPUS* ONLINE DE ESCRITA E FALA

Isabel Alves
Universidade de Lisboa. Faculdade de Letras / Centro de Linguística da Universidade de Lisboa
alvesisabel@campus.ul.pt
Patrícia Costa
Universidade de Lisboa. Faculdade de Letras
patriciacosta1@campus.ul.pt
Maria do Carmo Lourenço-Gomes
Instituto de Letras e Ciências Humanas - Universidade do Minho
mclgomes@ilch.uminho.pt
Celeste Rodrigues
Universidade de Lisboa. Faculdade de Letras / Centro de Linguística da Universidade de Lisboa
celesterodrigues@campus.ul.pt

Resumo

Neste artigo, apresentamos o funcionamento de um novo *corpus* de escrita e fala de crianças (Português LM) nos primeiros anos da escolaridade: a EFFE-On. O seu conteúdo está disponível *online* para pesquisas, dirigindo-se, principalmente, a professores, linguistas e terapeutas da fala. A EFFE-On foi desenvolvida no âmbito do projeto *Escreves como Falas - Falas como Escreves?* no Centro de Linguística da Universidade de Lisboa (CLUL). Entre outras aplicações, o *corpus* permite identificar formas ortográficas não-convencionais (FN-Cs); comparar ortografias não convencionais com as suas respetivas pronúncias; auxiliar o desenvolvimento e a adequação de materiais pedagógicos e terapêuticos; examinar aspetos fonológicos, morfológicos, sintáticos, discursivos, entre outros relativos ao texto produzido pela criança.

Palavras-chave

corpus online, ortografia, Educação, Linguística

Abstract

In this article, we present a new corpus of writing and speech of children (Portuguese L1) in the early years of schooling: EFFE-On. Its content is available online for research, addressing mainly teachers, linguists and speech therapists. The EFFE-On was developed within the project "Do you write as you speak - Do you speak as you write?" in the Linguistic Centre of the University of Lisbon. Among other applications, the corpus enables: the identification of unconventional spellings; the comparison of unconventional spellings and their corresponding pronunciations; the assistance of the development and adaptation of educational and therapeutic materials; the study of phonological, morphological, syntactic, discursive and other aspects related to the text produced by children.

Keywords

corpus online, spelling, Education, Linguistics

Résumé

Dans cet article, nous voulons présenter le fonctionnement d'un nouveau corpus de la langue écrite et de la langue parlée d'enfants (Portugais LM) aux premières années de scolarité: EFFE-On. Son contenu est disponible sur l'internet, destiné, essentiellement, aux enseignants, aux linguistes et aux orthophonistes. Le EFFE-On a été développé dans le cadre du projet «Écrivez-vous comme vous parlez - Parlez-vous comme vous écrivez?» au Centre de Linguistique de L'Université de Lisbonne. Parmi d'autres applications, le corpus permet: identifier orthographes non conventionnelles; comparer orthographes non conventionnelles avec leurs prononciations respectives; favoriser le développement et l'adaptation des matériels éducatifs et thérapeutiques; examiner des aspects phonologiques, morphologiques, syntaxiques, discursifs et d'autres concernant le texte produit par l'enfant.

Mots-clés

en ligne *corpus*, l'orthographe, l'Education, Linguistique

Resumen

En este artículo, presentamos el funcionamiento de un nuevo *corpus* de escritura y lenguaje infantil (con portugués como lengua materna) en los primeros años de escolaridad: EFFE-On. Su contenido puede ser consultado on-line y se dirige, principalmente, a profesores, lingüistas y terapeutas del lenguaje. EFFE-On fue desarrollado en el ámbito del proyecto "*Escreves como falas - falas como escreves?*" ("Escribes como Hablas-Hablas como Escribes?") en el *Centro de Linguística* de la Universidad de Lisboa. Entre otras aplicaciones, el corpus permite identificar formas ortográficas no convencionales, comparar ortografía no tradicional con sus respectivas pronunciaciones; ayudar en el desarrollo y adaptación de materiales pedagógicos y terapéuticos; examinando aspectos fonológicos, morfológicos, sintácticos y discursivos, entre otros, relativos a los textos producidos por los niños.

Palabras clave

corpus on-line, ortografía, Educación, Linguística

INTRODUÇÃO

A EFFE-On é uma plataforma *online*, criada no sistema TEI-TOK (Janssen 2014), que consiste num corpus de dados de escrita e fala de crianças nos primeiros anos de escolaridade, recolhidos transversal e longitudinalmente. Até ao momento, a plataforma possui textos de crianças do 2º e 4º anos de Lisboa (dados longitudinais) e do 2º ano do Porto.

Foi elaborado com base no projeto EFFE - Escreves como Falas - Falas como Escreves?[1], que começou a ser desenvolvido em 2012 no Centro de Linguística da Universidade de Lisboa (CLUL). O EFFE teve desde início dois objetivos distintos mas conciliáveis: por um lado, verificar se alguns erros ortográficos em textos do primeiro ciclo de escolaridade eram motivados pela fala e, por outro, reunir as produções orais e escritas que justificassem esses erros num banco de dados de fala e escrita correspondentes (Lourenço-Gomes, Rodrigues & Alves). Foi para cumprir este segundo desígnio que se criou a plataforma EFFE-On.

O principal objetivo deste *corpus* é fornecer estes dados a professores de Português língua materna, investigadores e terapeutas, possibilitando a fundamentação de um leque amplo de estudos em diferentes áreas, incluindo a Linguística, a Educação e a Terapia da Fala.

O conhecimento do sistema ortográfico pela criança é um processo gradual que se inicia no primeiro ano de escolaridade e se estende pelos anos seguintes. Sabendo que as crianças se encontram em diferentes momentos evolutivos, e que algumas enfrentam dificuldades, e em diferentes graus, é possível elaborar estratégias que respondam a essas diferenças. Aquelas crianças que se encontram em momentos mais maduros dessa evolução ajudam as outras nas suas dificuldades e, ao mesmo tempo, vão sedimentando conhecimentos já adquiridos. Por outras palavras, as atividades pedagógicas criadas com base numa verdadeira compreensão sobre o desenvolvimento individual e sobre o desenvolvimento de um determinado grupo beneficia todos porque otimiza o desenvolvimento de uns e minimiza as dificuldades de outros.

A preocupação com a formação do professor é constantemente ressaltada por diferentes autores. Goodman (1995) considera que o conhecimento do professor sobre a natureza da linguagem oral e escrita e sobre a natureza do desenvolvimento da criança habilita-a a escolher e a aplicar estratégias apropriadas a cada momento deste desenvolvimento, enfatizando que esta correlação é diretamente proporcional, isto é, quanto maior o conhecimento do professor, maior será esta capacidade. Zunino & Pizani (1995) acrescentam que conhecer o processo de aprendizagem, do ponto de vista do grupo e do ponto de vista de cada aluno, permite ao professor entender as produções e interpretações das crianças e, com isso, saber o que há por detrás das suas perguntas e respostas; determinar aquelas questões que podem tornar-se relevantes para elas naquele dado momento, antecipando as hipóteses das crianças a essas questões e, assim, elaborar estratégias pedagógicas apropriadas. Golbert (1988), por sua vez, enfatiza que a escola deve basear-se em conhecimentos teóricos essenciais sobre o desenvolvimento da criança e sobre a língua, para estar habilitada a ensinar. Acrescenta ainda que é importante conhecer as diferenças no que diz respeito a este desenvolvimento e à realidade sociocultural da criança, sem a considerar inferior ou menos capaz, uma vez que, apesar das diferenças, a grande maioria é potencialmente capaz de aprender a ler e a escrever.

A plataforma EFFE-On foi criada para pesquisas de diversa ordem e não apenas como ferramenta para contagem de frequências de Formas Não-Convencionais (daqui por diante, FN-Cs). As FN-Cs são entendidas neste artigo como formas originais produzidas pelas crianças que não obedecem às convenções ortográficas do Português. Não são assinaladas como FN-CS formas que, por razões sintáticas, de pontuação, ou outras, não obedeçam às convenções da escrita na Língua. A frequência de FN-Cs, isoladamente, não representa a maior ou menor dificuldade da criança durante o processo de aprendizagem da escrita. É importante compreender quais as características desses erros e o que representam em termos de construção do conhecimento. Também é preciso distinguir os erros que fazem parte dessa construção daqueles que evidenciam um conhecimento ainda elementar sobre o sistema de escrita como um todo por parte da criança. Muitas crianças no primeiro ano de escolaridade são capazes de perceber que não existe na nossa língua uma palavra grafada com dois "r" no início (como em *rraposa*), principalmente aquelas que estão familiarizadas com materiais escritos, porque foram expostas a eles mesmo antes de entrarem para o primeiro ano. O professor é quase sempre capaz de reconhecer as dificuldades dos alunos, mas nem sempre tem a segurança necessária acerca do seu significado. A presença dessas dificuldades pode ser um indicador de patologias várias que justifiquem o reencaminhamento para outro profissional (exceto em casos extremos, quando a escrita é tão repleta de alterações que se torna quase incompreensível, ou quando a criança apresenta dificuldades noutras áreas). Pode ser ainda difícil para o professor saber determinar que tipos de FN-Cs são esperados e os quais não são para uma determinada faixa etária. Também, e talvez não raramente, FN-Cs comuns são tratadas como *patológicas*

e FN-Cs atípicas como *normais*, dependendo de quem as cometeu. Lourenço-Gomes (1999) relata uma conversa que teve durante o seu estudo com uma professora do 2º ano do ensino básico:

Em uma reunião que tive com uma das professoras para apresentar os resultados da análise ortográfica, ela evidenciou reconhecer, entre os seus alunos, aqueles que apresentavam maiores dificuldades, aqueles que as apresentavam em menor grau e aqueles que ela considerava "muito bons". Mas quando mostrava, por exemplo, o número de alterações de um aluno classificado como "muito bom", ela ficava surpresa ("Mas ele é tão bom!"). Com a mesma surpresa ela recebia um comentário como "Muitos erros que ele apresenta são comuns às outras crianças", sobre um aluno classificado como "fraco". (Autor, 1999)

Talvez isso ocorra por desconhecimento sobre aspetos importantes do processo de aprendizagem da língua escrita e da própria língua. Sobre este tópico, no seu livro *Alfabetização e Linguística*, Cagliari (1989) ilustra a importância do conhecimento que o professor deve ter sobre a língua e sobre os processos de aquisição e desenvolvimento da linguagem oral e escrita.

Como é sabido, as investigações sobre a fala e a aprendizagem da leitura são muito mais numerosas do que aquelas sobre a aprendizagem da escrita. Perfetti (1997, p. 21), por exemplo, aponta que as razões para esta relativa negligência são múltiplas, mas incluem, pelo menos, o privilégio científico, herdado da Linguística, dado à linguagem falada. O autor acrescenta que a ortografia parece ser vista menos como um problema científico do uso da língua do que como uma convenção da alfabetização ou um assunto escolar. No final da década de 1990, surgem alguns estudos importantes que, mais diretamente, examinam os processos e a aquisição da ortografia em relação às estruturas linguísticas.

No entanto, esse cenário tende a modificar-se com o interesse crescente de linguistas e psicolinguistas sobre a natureza cognitiva da escrita (Cf. Guinet & Kandel, 2010; Shen, Damian & Stadthagen-Gonzalez, 2013; Kandel, Peereman & Ghimenton, 2014). Apesar de já na década de 80 se ter começado a observar algum interesse pelo estudo dos processos cognitivos da escrita (p. ex.,Frith, 1980) e de uma perspetiva mais próxima a esta abordagem aparecer em Treiman (1993), é no final da década de 90 que começam a surgir investigações que, mais diretamente, examinam os processos e a aquisição da ortografia relacionada com as estruturas linguísticas (Perfetti, Rieben & Fayol, 1997). Assim, o *corpus* que ora apresentamos pretende constituir uma importante ferramenta na recolha de contributos para o desenvolvimento da linguagem escrita em investigações com diferentes abordagens, sejam elas pedagógicas, linguísticas, psicolinguísticas ou clínicas.

Os dados do *corpus* foram recolhidos de acordo com procedimentos metodológicos sistemáticos e consistem em descrições de imagens com estímulos controlados e adequados às faixas etárias inquiridas. Estes procedimentos serão aplicados a qualquer recolha futura no âmbito da EFFE-On, que foi concebida para ser progressivamente ampliada. Até agora, o *corpus* contém produções, de natureza predominantemente descritiva, obtidas a partir de imagens-cenário para elicitação de um mesmo conjunto de palavras nas modalidades oral e escrita, escolhidas com base em critérios fonéticos, fonológicos, lexicais e de frequência (cf. Guerreiro, 2007: 99 - 113). Além disso, foram incluídas produções narrativas na componente escrita, construídas a partir de imagens em série – histórias sem texto (Furnari, 1993a e b).

No cenário da Língua Portuguesa, contendo uma análise mais descritiva sobre a ortografia, encontramos pelo menos três importantes grupos brasileiros que se dedicam a estudos sobre o desenvolvimento da escrita: GEALE, Didática da Língua Portuguesa e GPEL. O GEALE desenvolveu uma base de dados constituída por textos de crianças brasileiras e portuguesas nos primeiros anos de escolaridade e por textos produzidos por alunos do Programa EJA BRASIL (Educação de Jovens e Adultos). O Banco de Dado E-Labore já disponibiliza no sítio do grupo os materiais das recolhas realizadas na cidade de Belo Horizonte (MG). O volume especial de *Cadernos de Educação* (vol. 35, 2010) traz uma coletânea de investigações sobre a escrita infantil focando diferentes tópicos, sendo a maioria investigações em português brasileiro dos grupos de pesquisa acima mencionados.

METODOLOGIA

Participantes

Em Lisboa, na escola CM, foram recolhidos dados das mesmas crianças no 2º e 4º anos. No Porto, obteve-se, até ao momento, produções do 2º ano das escolas OsC e PA. As seguintes tabelas mostram as contagens dos participantes e das tarefas realizadas em cada uma das cidades.

Tabela 1: Participantes do *corpus* - dados de Lisboa

Ano e Turma	Crianças	Escolas	Sexos	Textos	Gravações	Idades
2A e 2B	48	CM privada	M-23 F-25	96	58	7
4A e 4B	54	CM privada	M-31 F-23	108	-	9

Tabela 2: Participantes do *corpus* - dados do Porto

Ano e Turma	Crianças	Escolas	Sexos	Textos	Gravações	Idades
2	44	OsC e PA pública	M-23 F-21	88	41	7

Materiais

Tarefa 1: Imagens em série (BR)

A primeira sessão de trabalho com os participantes do 2º ano tinha como propósito inicial familiarizá-los com o formato da atividade escrita posterior, funcionando como um estudo-piloto. Tinha-nos sido comunicado pelos professores que estes seriam os primeiros textos que as crianças iriam redigir autonomamente e sem conteúdo previamente fornecido. No entanto, a complexidade macrotextual e a riqueza vocabular dos dados veio a surpreender-nos. Por esse motivo, decidimos estender esta tarefa ao 4º ano e integrar todas as composições no material analisável e disponibilizá-las na plataforma EFFE-On, tornando possível a comparação de produções escritas do mesmo tipo das mesmas crianças, em duas fases distintas do seu percurso de escolarização.

Foram apresentadas às crianças vinhetas dispostas em série que, no seu conjunto, formam uma banda desenhada. Para o 2º ano, em Lisboa e no Porto, foi utilizada a história O Chapéu, extraída do livro A Bruxinha Atrapalhada (Furnari, 1993a). Para o 4º ano, escolheu-se uma outra menos infantil e mais adequada à faixa etária – O Telefone, retirada de O Amigo da Bruxinha (Furnari, 1993b).

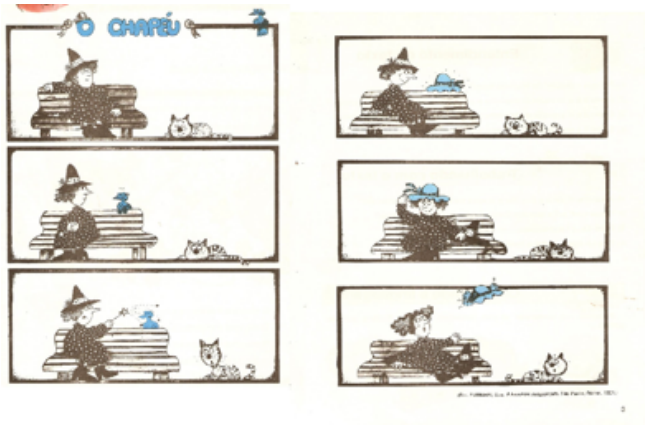

Figura 1: A Bruxinha Atrapalhada - O Chapéu (Furnari, 1993a)

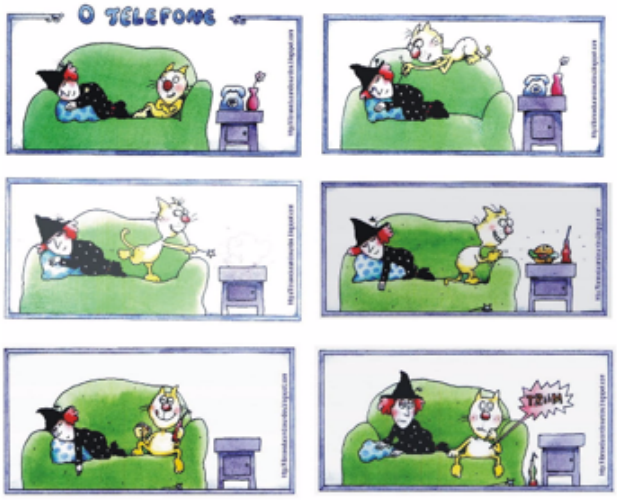

Figura 2: O Amigo da Bruxinha - O Telefone (Furnari, 1993b)

Tarefa 2: Imagens-cenário (FL, CI, SA, CB, CZ)

O principal objetivo era encontrar palavras-alvo ou estruturas comparáveis entre os textos e as produções orais, de modo a que se pudesse verificar a existência ou não de uma explicação de caráter oral para as FN-Cs (Formas Não-Convencionais).

Para obter essas produções, foram utilizadas cinco imagens-cenário - Floresta (FL), Cidade (CI), Sala (SA), Casa de Banho (CB) e Cozinha (CZ), cujo teor é predominantemente descritivo e narrativo. Cada uma contém elementos que servem como estímulo para a nomeação das palavras-alvo e para a descrição das cenas.

As figuras foram retiradas de um estudo anterior com propósitos distintos (Guerreiro, 2007). No entanto, escolheram-se estas tendo em conta que o instrumento continha, não só um conjunto de imagens-estímulo, como também um elenco definido de palavras-alvo, cujo equilíbrio fonético, fonológico, lexical e de frequência foi assegurado na conceção do material. Além disso, já tinha sido testado para fins linguísticos em duas variedades do Português (PE e PB), o que, no caso de se acrescentarem ao corpus dados do PB, poderá constituir uma mais-

-valia em futuros trabalhos de Linguística Comparada.
A uma das imagens - a Floresta (FL) - acrescentou-se um pequeno papel em dimensões aproximadas a A6 com a figura de um elefante. As crianças deveriam integrar o animal na situação narrada quer na fala, quer na escrita. Esta inovação ao suporte das imagens de Guerreiro (2007) prendeu-se com razões fonológicas, uma vez que se considerou importante testar uma estrutura que não constava no instrumento original e do tipo da presente no início desta palavra.
Pese embora o pré-teste aplicado tenha sido ajustado ao ano de escolaridade, as imagens da tarefa principal foram as mesmas nos 2º e 4º anos, de Lisboa e do Porto. Note-se que todas as redações obtidas nesta tarefa foram incluídas na EFFE-On.

Figura 3: Imagens-Cenário

Procedimentos de recolha de dados

Para recolha e uso dos dados, foram requeridas autorizações e preenchidos questionários pelos encarregados de educação de cada informante para a obtenção de informações fundamentais definidoras do perfil de cada criança, como as línguas com que contactam e eventuais dificuldades na aprendizagem (informações a disponibilizar na plataforma).
Num primeiro dia, os informantes foram confrontados com o pré-teste e, num segundo, com a tarefa principal - de manhã o exercício escrito, de tarde o exercício oral, sempre acompanhados de um exemplar em formato A4 da imagem a ser descrita. Para esta secção foram despendidas cerca de duas horas, incluindo o tempo de interação do investigador com as crianças. Importa salientar que as orientações fornecidas foram integralmente gravadas de maneira a garantir um relatório metodológico fidedigno e pormenorizado.
Os participantes redigiram ambos os textos em folha pautada A4, em turma, com a presença do professor e do investigador - que apresentou e descreveu as imagens, salientando as palavras-alvo. As crianças foram instruídas para que as composições concentrassem os géneros descritivo e narrativo. Se a descrição potencia o detalhe, a narração promove a criatividade, motivando palavras não diretamente evocáveis pela figura, todavia potencialmente interessantes, como sejam as formas verbais.
Note-se que as cinco imagens temáticas foram distribuídas pelas turmas/ anos de forma equitativa, de modo a que se obtivesse aproximadamente o mesmo número de produções de cada figura e que cada criança se debruçasse apenas sobre uma mesma imagem, na escrita e na oralidade.
No que diz respeito à tarefa de oralidade, realizaram-se sessões individuais gravadas na íntegra com uma duração média de 10 minutos. Para o efeito, reservou-se uma sala da escola, previamente visitada pelo investigador, reunindo as condições necessárias - o mínimo de qualidade acústica e o conforto da criança. No compartimento escolhido, estavam presentes apenas a criança e o investigador. Em primeiro lugar, foi pedido a cada participante que recontasse a história que tinha redigido para que coincidissem o mais possível os vocábulos da fala com os da escrita. Em seguida, incentivou-se a livre descrição da figura e da situação em causa e a nomeação dos objetos e das respetivas cores, para abranger o máximo de elementos presentes na imagem.
Durante toda a entrevista, o investigador tinha consigo um guião com algumas questões norteadoras do diálogo e uma lista das palavras-alvo e ia registando a produção de cada uma, de forma a que todas fossem mencionadas.

Tabela 3: Lista das Palavras-Alvo

Fonte	Imagem	SA	CI	CB	FL	CZ
Guerreiro (2007)		Almofada Caixa Dinheiro Desenho Lápis Jornal Tesoura Rádio Televisão Quadro Igreja Bruxa	Carro Grande Andar Bicicleta Estrada Futebol	Menina Espelho Escova Torneira Toalha Orelha Pescoço Botão	Nuvem Sol Trovoada Voar Dragão Fogo Floresta Verde Peixe Borboleta Caracol Flor	Estrela Fogão Gelado Comer Rato
Acrescentadas em todas as turmas		Prateleira Voar Vassoura	Vermelho Piscina Touca Telhas	Espuma Transbordar Fechar Chave	Elefante	Queijo Parabéns
Acrescentadas em todas as turmas, exceto 2B_CM			Chaminé Fumo	Esponja Banheira Azulejos		

FUNCIONAMENTO DA BASE DE DADOS

A plataforma tem como suporte o sistema TEITOK, concebido para anotar, disponibilizar e visualizar *corpora online* com formato TEI - Text-Encoding Initiative (Janssen, 2014). Até ao momento, no CLUL, o TEITOK associou-se aos projetos *P.S.: Post-Scriptum*, COPLE2: *Learner Corpus of Portuguese L2* e, mais recentemente, *EFFE-On: Escreves como Falas, Falas como Escreves? - Online*.

O acesso à EFFE-On[2] possibilita a pesquisa e visualização de materiais de diferentes tipos, designadamente, imagens fac-similadas dos textos originais, ficheiros-áudio, transliterações codificadas em XML que permitem confrontar a versão original da criança com versões com formas normalizadas das FN-Cs e sem segmentos riscados, ilegíveis ou acrescentados fora da linha. O resultado das pesquisas efetuadas pode ser descarregado num ficheiro com formato TXT. A equipa da EFFE-On disponibiliza os seus contactos para o esclarecimento de qualquer dificuldade ou dúvida relativamente à utilização da plataforma.

Identificação dos ficheiros

Cada ficheiro presente na EFFE-On está devidamente identificado com um código, que contém os seguintes elementos:

- número do participante (atribuído aquando da análise dos dados);
- abreviatura da designação do estímulo/tarefa;
- ano e turma do participante;
- abreviatura da designação da escola onde foi efetuada a recolha.

Assim, um texto redigido pelo participante 10, acerca da descrição da imagem-cenário Sala, da turma B do 2.º ano e da escola CM, tem a seguinte identificação: 10_SA_2B_CM. Por defeito, o nome do ficheiro não compreende a referência à localização geográfica das escolas onde foram efetuadas as recolhas. No entanto, para distinguir os dados do Porto (e os dados que futuramente enriquecerem o corpus, provenientes de outras localizações), convencionou-se que os códigos devem incluir essa informação. Deste modo, um texto produzido pelo participante 70, acerca da história A Bruxinha Atrapalhada - O Chapéu, do 2.º ano da escola CT do Porto, tem a seguinte identificação: P_70_BR_2_CT. Em todo o caso, os materiais disponíveis da região de Lisboa apresentam no respetivo cabeçalho a localização da escola em que foram recolhidos (CM – Lisboa).

Opções de visualização dos dados

O *corpus* é atualmente constituído por 292 produções escritas realizadas por 146 alunos, do 2.º e 4.º anos de uma escola particular em Lisboa e do 2.º ano de duas escolas públicas no Porto. Para além dos textos, a plataforma é constituída por 99 ficheiros áudio associados.

Cada ficheiro disponibilizado online é constituído pelos seguintes itens:

(i) o código de identificação;
(ii) o título da tarefa;
(iii) o cabeçalho com informações acerca do participante;
(iv) a transliteração do texto original produzido pela criança;
(v) a imagem fac-similada do texto original;
(vi) o ficheiro áudio com produções orais gravadas.

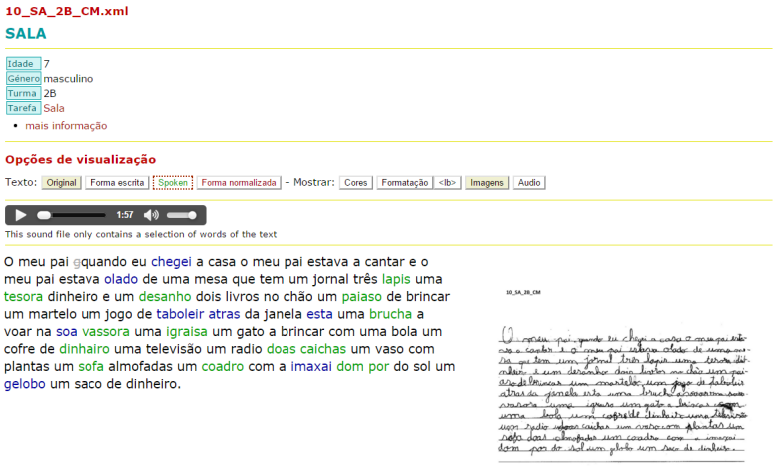

Figura 4: Interface da EFFE-On

O cabeçalho de cada ficheiro contém informações pessoais acerca do participante (idade, género, data de nascimento, nacionalidade, histórico linguístico, histórico clínico de patologias da fala, ano e turma e escola), informações relativas ao grau de escolaridade dos pais da criança e informações respeitantes à identificação do ficheiro (tipo de tarefa, áudio). Apenas quatro destes campos estão visíveis (idade, género, turma e tarefa). A consulta da versão integral do cabeçalho faz-se clicando em mais informação.

O texto transliterado resultante da produção original da criança é apresentado ao utilizador, conservando títulos, parágrafos, segmentos riscados, ilegíveis ou acrescentados acima da linha, bem como Formas Não-Convencionais. Estas ocorrências, codificadas em formato XML, são assinaladas com cores diferentes, facilitando a sua distinção do restante conteúdo textual. A edição linguística do texto permite ao usuário confrontar as formas originalmente escritas com as contrapartidas normalizadas e orais.

As restantes opções de visualização modificam o texto, resultando em três variantes do original: a seleção de forma escrita omite os segmentos riscados e ilegíveis, mantendo apenas a versão definitiva da criança; a opção spoken substitui as FN-Cs pelas formas orais correspondentes e a opção forma normalizada substitui as formas originais não-convencionais pelas suas contrapartidas convencionadas.

A plataforma dispõe ainda da possibilidade de visualizar ou omitir a imagem fac-similada que corresponde ao texto transliterado (Imagens).

Os ficheiros áudio são constituídos por uma seleção de palavras produzidas oralmente com correspondência no texto escrito. Não disponibilizamos os áudios integralmente porque as entrevistas não compreendem uma leitura dos textos escritos produzidos por cada uma das crianças, mas descrições mais espontâneas feitas oralmente. Repare-se que o ficheiro áudio apresenta as palavras produzidas separadas por intervalos de silêncio com 8ms de duração, de modo a permitir a concentração em cada nova palavra. Sempre que foi considerado pertinente, adicionou-se, a cada uma, o contexto linguístico de ocorrência oral (que não coincide obrigatoriamente com o contexto escrito).

Opções de pesquisa dos dados

A extração de dados na EFFE-On segue o CQP – Corpus Query Protocol –, que possibilita a pesquisa simples no texto por palavra, sequência de palavras, lema ou etiquetas POS (Part-of-Speech), fornecendo listas de ocorrências em vários contextos e listas de frequência.

A pesquisa avançada, por seu turno, permite ao utilizador restringir e orientar os resultados obtidos. Pode, com esta opção, procurar no corpo do texto ou nos ficheiros. Os campos pesquisáveis no texto são os mesmos que na pesquisa simples enquanto que, nos ficheiros, as opções de procura são as seguintes: identificação da criança, idade, género, ano/turma, escola, escolaridade dos pais e tarefa. A exploração deste tipo de variáveis é especialmente importante para investigadores, professores, terapeutas e demais utilizadores da plataforma que pretendam traçar correlações entre os dados linguísticos e aspetos sociolinguísticos.

Figura 5: Pesquisa avançada na EFFE-On

APLICAÇÕES

A EFFE-On permite aos utilizadores o acesso livre aos materiais disponibilizados e editados. Através deles, linguistas, professores de língua materna ou terapeutas da fala poderão formular hipóteses acerca da relação entre escrita e fala - nomeadamente, no que concerne à influência da oralidade na produção de FN-Cs - ou concentrar-se numa só dimensão discursiva, escrita ou fala.

Assim, entre outras aplicações, o corpus permitirá:

(i) identificar FN-Cs, em textos produzidos por crianças falantes do Português Europeu de, pelo menos, duas variedades dialetais[3];

(ii) confrontar as FN-Cs e as produções orais espontâneas da mesma criança e entre crianças;

(iii) verificar se, de facto, existe uma correlação entre as FN-Cs e as respetivas formas na oralidade;

(iv) comparar dados recolhidos num determinado momento (recolhas transversais) e ao longo do tempo (recolhas longitudinais);

(v) auxiliar o desenvolvimento e adequação de materiais pedagógicos e terapêuticos, reconhecendo as diferentes características da variedade falada em determinada região;

(vi) descrever e interpretar aspetos fonológicos, morfológicos, sintáticos, discursivos, entre outros, relativos ao texto e

(vii) pesquisar frequências de palavras e de estruturas sintáticas.

CONSIDERAÇÕES FINAIS

Neste artigo procurámos apresentar e descrever o funcionamento da nova base de dados de escrita e fala nos primeiros anos de escolaridade do ensino básico. Esta plataforma, criada no âmbito do projeto EFFE - Escreves como Falas, Falas como Escreves? - destina-se a todos os profissionais interessados na interpretação e descrição dos dados de escrita relacionados com a oralidade. Para além das aplicações mencionadas na secção anterior, outras potencialidades poderão vir a ser estudadas em função dos propósitos dos diferentes trabalhos.

Até ao final de 2015, a EFFE-On será disponibilizada *online* para consulta dos dados de Lisboa e, no início de 2016, nas suas completas valências juntamente com os dados do Porto (com a opção de pesquisas avançadas e com a totalidade dos ficheiros áudio).

Como avançado na Introdução, a equipa do EFFE propôs-se, até ao momento, a atingir dois intuitos distintos. Por um lado, num *corpus* transversal e longitudinal, elencar dados de várias cidades portuguesas e anos de escolaridade. Esse objetivo começou a ser cumprido com a EFFE-On, todavia, no ano de 2016, afigura-se possível a extensão das recolhas a novas regiões e variedades do Português. Por outro lado, o grupo planeia explorar, através dos dados de correspondência entre fala e escrita reunidos, as Formas Não-Convencionais encontradas nos textos, enveredando por estudos de índole fonológica e psicolinguística. Para tal, o EFFE pretende elaborar uma interpretação fonológica dos dados de escrita relacionados com a oralidade, além de explorar esses dados em experiências psicolinguísticas que possibilitem, através de técnicas *online*, compreender a natureza cognitiva do conhecimento ortográfico (Cf. Lourenço-Gomes, Rodrigues & Alves).

Notas

1 Financiado por FCT, projeto UID/LIN/002114/2013

2 Através do endereço http://alfclul.clul.ul.pt/teitok/effe/

3 Para tal, o EFFE pretende elaborar uma interpretação fonológica dos dados de escrita relacionados com a oralidade, além de explorar esses dados em experiências psicolinguísticas que possibilitem, através de técnicas *online*, compreender a natureza cognitiva do conhecimento ortográfico (Lourenço-Gomes, Rodrigues & Alves)

Referências

Cagliari, L.C. (1989). *Alfabetização e linguística*. Rio de Janeiro: Scipione.

Frith, U. (Edit.). (1980). *Cognitive Process in Spelling*. London: Academic Press.

Golbert, C. (1988). *A evolução psicolinguística e suas implicações na alfabetização: teoria, avaliação, reflexões*. Porto Alegre: Artes Médicas.

Goodman, Y. M. (Org.). (1995). Conhecimento das crianças sobre a alfabetização: um posfácio. In Y. M. Goodman, *Como as crianças constroem a leitura e a escrita*: perspectivas piagetianas. Porto Alegre: Artes Médicas.

Guerreiro, H. (2007). *Processos fonológicos na fala da criança de cinco anos*. Tese de Mestrado. Instituto de Ciências da Saúde da Universidade Católica Portuguesa-Escola Superior de Saúde do Alcoitão da Santa Casa da Misericórdia de Lisboa.

Guinet, E. & Kandel, S. (2010). Ductus: a software package for the study of handwriting production. *Behavior Research Methods*, 42 (1), 326-332.

Furnari, E. (1993a). *A bruxinha trapalhada*. São Paulo: Global.

Furnari, E. (1993b). *O amigo da bruxinha*. São Paulo: Moderna.

Kandel, S., Peereman, R. & Ghimenton, A. (2014). How do we code the letters of a word when we have to write it? Investigating double letter representation in French. *Acta Psychologica*, 148, 56-62.

Lourenço-Gomes, M. C. (1999). *Leitura e Escrita e Consciência Fonológica*: intervenção em sala de aula. Monografia (Especialização em Distúrbios da Comunicação Humana). Universidade Federal do Estado de São Paulo/ Universidade Católica de Petrópolis.

Lourenço-Gomes, M. C., Rodrigues, C. & Alves, I. (no prelo). Escreves como Falas - Falas como escreves? *Revue Romane.*

Perfetti, C. A. (1997).The Psycholinguistics of spelling and reading. In C. A. Perfetti, L. Rieben & M. Fayol (Eds.), *Learning to spell: research, theory, and practice across languages* (pp. 21-38). NJ: Lawrence Erlbaum.

Perfetti, C. A., Rieben, L. & Fayol, M. (1997) *Learning to spell: research, theory, and practice*. NJ: Lawrence Erlbaum.

Shen, X. R., Damian, M. F. & Stadthagen-Gonzalez, H. (2013). Abstract graphemic representations support preparation of handwritten responses. *Journal of Memory and Language,* 68 (2), 69-84.

Treiman, R. (1993). Beginning to spell: A study of first-grade children. NY: Oxford University Press.

Yavas, M., Hernandorena, C. L. M & Lamprecht, R. R. (1991). *Avaliação fonológica da criança*. Porto Alegre: Artes Médicas.

Zunino, D. L. & Pizani, A. P. (1995). *A aprendizagem da língua escrita na escola: reflexões sobre a proposta construtivista* (2ª ed.). Porto Alegre: Artes Médicas.