LexPorBr Infantil: uma base lexical tripartida e com interface Web de textos ouvidos, produzidos, e lidos por crianças

Gustavo Estivalet[1], Nathan S. Hartmann[2], Vanessa Marquiafável[3], Katerina Lukasova[4], Maria T. Carthery-Goulart[4], Sandra M. Aluísio[2]

[1]Federal University of Paraíba, Department of Classical and Vernacular Letters

[2]University of São Paulo, Institute of Mathematics and Computer Sciences

[3]SpeechTera Desenvolvimento de Programas para Computadores Ltda

[4]Federal University of ABC, Center for Mathematics, Computation and Cognition

gustavoestivalet@hotmail.com

{nathansh, sandra}@icmc.usp.br

marquiafavel@gmail.com

{teresa.carthery, katerina.lukasova}@ufabc.edu.br

Abstract. *Corpora of words have been widely used in the selection of stimuli in psycholinguistic experiments, research in lexicology, among others uses. Word frequency is a great proxy in psycholinguistics research, since it provides a good time response in word recognition. This paper presents the first of the three corpora of the LexPorBr Infantil project: subtitles of family, comedy, and children movies and series listen by children in Brazilian Portuguese. This work provides a large lexicon of words, publicly available, with almost 130 million tokens and 880 thousand types where each word is annotated with 48 categories for psycholinguistic research, corpus analysis and research in education.*

Resumo. *Córpus de palavras têm sido largamente utilizados na seleção de estímulos em experimentos psicolinguísticos, pesquisas em lexicologia, dentre outros usos. A frequência de palavras é um importante proxy em pesquisas psicolinguísticas, pois prevê com boa precisão os tempos de reação para o reconhecimento de palavras. Este artigo apresenta o primeiro dos três córpus do projeto LexPorBR Infantil: legendas de filmes e séries de comédia, família e animações em português brasileiro ouvidos por crianças. Este trabalho disponibiliza publicamente um léxico de 130 milhões de tokens e 880 mil types, disponibilizando 48 categorias de informações para pesquisas em psicolinguística, análise de córpus e aplicadas à educação.*

1. Introdução

Pesquisas em linguística de córpus têm sido uma poderosa ferramenta para o estudo das línguas naturais nos mais diversos níveis de análise assim como na interface com várias disciplinas. Podemos diferenciar córpus de textos baseados na análise de sentenças e córpus de palavras baseados na análise de entradas lexicais [Brysbaert and New 2009].

A frequência de palavras é uma das variáveis mais importantes nas pesquisas em psicolinguística, estabelecendo que palavras que são lidas e ouvidas mais frequentemente são reconhecidas mais rapidamente do que palavras menos frequentes [Brysbaert et al. 2011a]. Os córpus a partir dos quais as frequências das palavras são calculadas devem conter entradas variadas e representativas da língua (ouvida e/ou escrita). Além disso, é interessante que estes córpus possam ser ajustados para populações específicas, uma vez que a exposição a conteúdos linguísticos varia conforme a idade, nível socioeconômico, grau de escolarização, entre outros.

Em geral, os córpus são compostos por um grande conjunto de livros, jornais e revistas que têm oferecido parâmetros satisfatórios para a população adulta e escolarizada, apesar de alguns contrastes em relação à familiaridade de entradas lexicais cuja ocorrência predomina na linguagem oral ou na escrita. Do ponto de vista lexical, a linguagem escrita tende a ser mais redundante em função da falta de interação comunicativa, o que pressupõe a escolha de itens lexicais menos ambíguos. Assim, córpus construídos unicamente a partir de textos escritos são mais problemáticos para estudos psicolinguísticos envolvendo crianças.

Para lidar com estes vieses, pesquisas em linguística de córpus têm incluído materiais que possam se aproximar da linguagem oral para maior precisão na obtenção de parâmetros de frequência. Nesse sentido, mensagens de texto, e-mails e legendas de filmes e programas de televisão têm sido apontados como materiais mais representativos da língua falada. Alguns estudos mostraram que a frequência obtida a partir de legendas prevê com mais precisão os tempos de reação para o reconhecimento de palavras [Brysbaert and New 2009, Brysbaert et al. 2011a, Brysbaert et al. 2011b, van Heuven et al. 2014].

Uma vantagem adicional do uso de legendas é a possibilidade de se ajustar o córpus para finalidades específicas a partir de filtros, por exemplo, estudos com crianças de diferentes faixas etárias e escolaridades. O objetivo das bases lexicais infantis é oferecer uma ferramenta sensível para estudos psicolinguísticos e de desenvolvimento com controle das variáveis psicolinguísticas para cada faixa etária [Corral et al. 2009]. Para tais propósitos, o uso de córpus baseado em textos escritos que somente a população adulta está exposta tem recebido críticas, pois as variáveis psicolinguísticas, tais como frequência das palavras, tendem a não refletir a realidade linguística do mundo infantil, uma tendência observada nos córpus já existentes. Assim, o crescente interesse em desenvolver estudos para crianças com foco experimental requer a existência de bases de dados lexicais contendo informação psicolinguística ajustada para esse público.

Com o objetivo de se responder a seguinte pergunta: *quais são as palavras que as crianças escutam com maior frequência*, este artigo apresenta o primeiro dos três córpus do projeto LexPorBR Infantil, que inclui: i) legendas de filmes e séries ouvidos por crianças, ii) textos escritos por crianças e iii) textos lidos por crianças. O LexPorBR Infantil - Oral (Legendas) foi compilado a partir de filmes e séries classificados de acordo com o córpus SubIMDb-PT [Paetzold and Specia 2016] como gêneros familiares (filmes/séries para família, comédia, crianças e animação), pois esses gêneros são destinados a crianças, famílias e/ou para todos, apresentando assim uma linguagem

acessível[1]. Legendas de filmes e séries têm sido consideradas material linguístico que apresentam o vocabulário mais próximo do dia-a-dia e da linguagem oral, logo, possuem normas de frequência otimizadas para a formação de córpus no estudo sobre a aquisição, o processamento e a utilização da linguagem tanto por crianças como por adultos [Brysbaert and New 2009, Soares et al. 2014a].

Este trabalho teve dois objetivos principais: i) disponibilizar de forma pública um grande léxico de palavras, com 129.053.297 *tokens* e 874.887 *types*, acessado por uma interface Web; ii) calcular 48 categorias importantes para pesquisas em psicolinguística, análise de córpus e aplicadas à educação (por exemplo: lexema, forma fonológica, POS, flexão nominal (gênero e número) e flexão verbal (modo, tempo, pessoa e número), silabação, sílaba tônica, número de letras, fonemas e sílabas, vizinhos ortográficos e fonológicos, OLD20/PLD20 (Orthographic/Phonological Levenshtein Distance).

2. Trabalhos Relacionados

Lexin [Corral et al. 2009] é uma base de vocabulário infantil do Espanhol composta por 134 livros destinados a estimular a leitura e a escrita em crianças na pré-escola (76 livros) e no primeiro ano do ensino fundamental (58 livros). A base contém 13.184 palavras (*types*) e 178.839 *tokens*. Manulex [Lété et al. 2004] é uma base do Francês composta por textos da 1ª a 5ª série da escola fundamental (idade 6 a 11 anos). A seleção do material foi feita considerando-se a frequência cumulativa referente aos livros mais vendidos pelas principais editoras francesas no ano de 1996. A base contém um total de 48.886 palavras e 23.812 lemas. Novlex [Lambert and Chesnet 2001] é mais uma base do Francês que descreve material didático da 3ª série e de leitura correspondente (19 títulos). A base totaliza 20.600 palavras e 9.300 lemas.

Entre as variáveis que são geralmente computadas nos referidos córpus, encontram-se: frequência das palavras, ocorrência de uma palavra em diferentes textos, frequência por mil/milhão, log da frequência por mil/milhão, categoria gramatical, número de letras, estrutura silábica, ano escolar com maior probabilidade da criança se deparar com a palavra, além de outras variáveis relacionadas com as estrutura ortográfica e fonológicas das palavras.

Em português europeu, Escolex [Soares et al. 2014b] é um córpus composto por 3,2 milhões de palavras (3.211.805 *tokens* e 48.381 *types*) coletadas de uma base de 171 livros escolares do 1º ao 6º ano do Ensino Fundamental (crianças de 6 a 11 anos). Além das medidas já relatadas nas outras bases, Escolex estima também diversidade contextual.

Em português brasileiro, o projeto que mais se aproxima do LexPorBR Infantil - Oral (Legendas) é o SUBTLEX-PT-BR [Tang 2012], compilado a partir do site de legendas OpenSubtitles em dezembro de 2012, com 61 milhões de *tokens* e 136.147 *types*, mas que não traz conteúdo direcionado para crianças. Este córpus disponibiliza: i) unigramas com OLD20; ii) lemas e POS *tagging*; e iii) bigramas, que são úteis para obtenção da frequência das colocações e para a identificação de palavras compostas.

[1]Entretanto, cabe aqui uma ressalva: o córpus foi compilado automaticamente e assim incluiu filmes com temas adultos do gênero comédia, em alguns casos.

3. O Processamento do córpus LexPorBR Infantil - Oral (Legendas)

O SubIMDb-PT é um córpus composto por legendas de filmes e séries infantis, utilizado no trabalho de [dos Santos et al. 2017], mas não descrito anteriormente. SubIMDb-PT foi compilado a partir do site de legendas OpenSubtitles em janeiro de 2017 e foi utilizado para o desenvolvimento do presente trabalho, o LexPorBR Infantil - Oral (Legendas), seguindo a mesma metodologia apresentada para a compilação do SubIMDb-EN [Paetzold and Specia 2016]. O córpus base não contém marcações que distinguem o início do fim da legenda de cada produção, pois cada filme/série dos gêneros familiares em português brasileiro foi concatenado em um único arquivo, sem anotação.

3.1. Segmentação, limpeza e tokenização das legendas

O gênero legendas é caracterizado por sentenças curtas, pois as mesmas devem ser dispostas na tela da televisão. Muitas sentenças, no entanto, não são suficientemente curtas e parte delas é exibida em diferentes *frames* na televisão. No córpus SubIMDb-PT, esse fenômeno é representado por sentenças quebradas por orações, seguidas por quebras de linhas. Para lidar com esse fenômeno, foi necessária a reconstrução das sentenças do córpus. Ainda, foram aplicados filtros de remoção de: (i) marcações de fala (travessão), tornando o córpus de legendas em um texto corrido; (ii) endereços de sites; (iii) referências ao editor/criador da legenda (conteúdo recorrente em legendas traduzidas autores amadores); (iv) *tokens* contendo caracteres diferentes do alfabeto do português brasileiro. Após o pré-processamento inicial, o córpus foi tokenizado com o *TreebankWordTokenizer* do NLTK[2]. A Tabela 1 apresenta estatísticas em relação aos *tokens*, *types* e *type/token ratio* (TTR) identificados durante o pré-processamento, assim como estatísticas de dois projetos relacionados ao LexPorBR Infantil - Oral (Legendas).

Córpus	**Tokens**	**Types**	**TTR**
LexPorBR Infantil - Oral (Legendas) original	168.888.430	927.023	0,55%
Aplicação de filtros de limpeza realizados	129.053.297	874.887	0,68%
Aplicação de Léxico de língua UNITEX (DELAF)	121.281.557	289.001	0,24%
Escolex	3.211.805	48.381	1.50%
SUBTLEX-PT-BR	61.000.000	136.147	0,22%

Tabela 1. Distribuição de *tokens*, *types* e *type/token ratio* do LexPorBR Infantil.

Com base no TTR, percebemos que o córpus possui maior riqueza lexical do que o SUBTLEX-PT-BR e menor riqueza lexical do que o Escolex. Observa-se ainda que, após a aplicação dos filtros, houve um aumento da riqueza lexical tendo em vista que houve, proporcionalmente, uma grande diminuição de *tokens* e uma pequena diminuição de *types*. Quando comparado com o léxico do dicionário UNITEX-PB DELAF [Muniz 2004], houve uma diminuição da riqueza lexical em função da grande diminuição de *types* de baixa frequência e, consequentemente, baixa diminuição de *tokens*.

Calculamos também a distribuição de frequência de *types* por decil, onde 50% deles possuem uma única ocorrência (cauda longa), 10% possuem 2 ocorrências, 10% possuem 3 ou 4 ocorrências, 10% possuem de 5 a 7 ocorrências, 10% possuem entre 8 e 24 ocorrências e somente 10% possuem mais que 25 ocorrências, podendo chegar a

[2]`https://www.nltk.org`

frequência máxima de 3.480.284. A baixa frequência para a grande maioria dos *types* do córpus pode ser justificada por erros de digitação, problemas de codificação do arquivo original da legenda (produzindo caracteres estranhos) e, claro, palavras raras para o gênero. Exemplos de *types* com frequência 1 no córpus são: N'attend, Novo.então, Peregrinaã§ães, Opelette, L-e-v-e-d-a-ç-ã-o, Estiércol, Pórticos, Gabiente, Ruptured.

Os *types* com frequência 1 foram avaliados no DELAF, sendo que apenas 33% deles são palavras da língua. Essa mesma análise para todo o córpus (linha 3 na Tabela 1) implica na redução de 68,8% no número de *types*, indicando o alto número de *tokens* ruidosos ou não presentes no DELAF. No entanto, vale lembrar que este dicionário não contempla uma grande variedade de nomes próprios, neologismos e estrangeirismos. Assim, com o objetivo de preservar palavras raras e nomes próprios, não removemos *types* de baixa frequência.

3.2. Tagging e Lematização

Para realizar o *POS tagging* e lematização no LexPorBR Infantil - Oral (Legendas), utilizamos o nlpnet [Fonseca et al. 2015] alinhado com o dicionário UNITEX-PB DELAF para mapearmos os lexemas com suas respectivas etiquetas morfossintáticas para o lema adequado. O nlpnet é um etiquetador morfossintático amplamente utilizado, treinado no córpus MacMorpho [Aluísio et al. 2003], que foi revisado para melhorar a tarefa de *POS tagging*[3], mas que não possui o mesmo conjunto de etiquetas morfossintáticas do DELAF. Para realizar o mapeamento entre as categorias do DELAF e as 25 categorias morfossintáticas da versão 3 do MacMorpho utilizada no nlpnet, fizemos um relaxamento nas etiquetas do DELAF, considerando somente as categoria principal e geral.

3.3. Outros léxicos utilizados no estudo

A fim de estudarmos a aderência do vocabulário utilizado nas legendas do nosso córpus com o esperado por crianças, fizemos uso de dicionários sugeridos pelo Programa Nacional do Livro Didático (PNLD) do Ministério da Educação (MEC). Esses dicionários foram categorizados por níveis de complexidade lexical esperada em cada etapa escolar, previamente compilados no trabalho de [Hartmann et al. 2018]. O dicionário de Tipo 1, composto aqui pelo dicionário Caldas Aulete com a Turma do Cocoricó, contempla o 1º ciclo do Ensino Fundamental 1 (1º ao 3º ano) e possui 1.371 entradas; o dicionário de Tipo 2, composto pelo Dicionário Escolar da Língua Portuguesa, Dicionario Ilustrado de Português e Dicionário Escolar da Língua Portuguesa Ilustrado com a Turma do Sítio do Pica-Pau Amarelo, contempla o 2º ciclo do Ensino Fundamental 1 (4º ao 5º ano) e possui 8.171 entradas; e o dicionário de Tipo 3, composto aqui pelo Minidicionário Contemporâneo da Língua Portuguesa, contempla o Ensino Fundamental 2 (6º ao 9º ano) e possui 29.970 entradas.

Também utilizamos léxicos amplamente utilizados em pesquisas do Processamento de Linguagem Natural: o UNITEX-PB, já apresentado nessa seção, contendo 7.580.357 palavras; e o Hunspell, dicionário eletrônico frequentemente utilizado em recursos computacionais, contendo 312.418 palavras.

[3] `http://nilc.icmc.usp.br/macmorpho`

4. O transcritor fonético Petrus e as categorias de silabação, sílaba tônica e transcrição fonológica

O processo de conversão de textos ortográficos em seus correlatos sonoros é chamado de conversão grafema-fonema (do inglês *grapheme-to-phoneme* - G2P) ou transcrição letra-som. O Petrus 2.0 (*PhonEtic TRanscriber for User Support*) [Serrani 2015] foi o sistema de conversão G2P utilizado para a obtenção da silabação, da indicação do acento primário e da transcrição fonológica das palavras contidas no LexPorBR Infantil - Oral (Legendas).

4.1. Divisão silábica e sílaba tônica

A metodologia adotada para a marcação da sílaba tônica (acento primário) em palavras simples do português brasileiro foi baseada nas regras publicadas por [Silva et al. 2006] devido à completa documentação e disponibilização dos algoritmos desenvolvidos. A taxa de acerto de 93% obtida em um córpus de teste composto por 52.525 palavras também foi decisiva para sua escolha, visto ser a maior entre os sistemas testados.

Vale mencionar que existem diferenças entre o conjunto de regras adotados para uma divisão silábica para efeitos de translineação e uma divisão silábica feita com base fonológica. Os algoritmos de silabificação propostos por [Silva 2011] foram desenvolvidos com a intenção de conciliar as teorias fonológicas da língua com as necessidades de sistema de síntese de fala. A Tabela 2 traz exemplos de palavras divididas silabicamente conforme as regras adotadas pelo Petrus e pelo Dicionário online Caldas Aulete.

	Petrus	**Dicionário online Caldas Aulete**
obstrução	o.bs.tru.ção	obs.tru.ção
advogado	a.d.vo.ga.do	ad.vo.ga.do
arredondar	a.rre.don.dar	ar.re.don.dar
assado	a.ssa.do	as.sa.do

Tabela 2. Divisão silábica do Petrus e do Dicionário online Caldas Aulete.

4.2. Transcrição fonológica

A obtenção da transcrição fonológica se deu da seguinte forma: o último módulo do referido sistema é responsável por realizar a transcrição fonética das palavras, que é feita a partir de diferentes níveis de informação sobre a palavra em análise, desde a presença ou não de um prefixo até a categorização gramatical, identificação da vogal tônica e da divisão silábica. Dessa forma, criou-se manualmente um conjunto de regras linguísticas dependentes de contexto (silábico, acentual e gramatical), que aliado ao uso de um dicionário fonético (ou seja, uma lista de palavras cujas transcrições fonéticas não seguem as regras de transcrição propostas) indica como transcrever os grafemas em suas respectivas unidades fonéticas. Os resultados obtidos com o Petrus indicaram uma taxa de acerto de 97.5% ao fone [Serrani 2015].

Por fim, a partir do *output* fonético gerado ao fim do processamento, e com base em uma lista de correspondência fone-fonema, fez-se a conversão da transcrição fonética gerada em IPA para uma transcrição fonêmica em alfabeto SAMPA adaptado. O Petrus foi desenvolvido para transcrever palavras simples do português brasileiro. Portanto, palavras compostas, estrangeirismos que não estejam em sua base de dados, palavras com

erros ortográficos ou qualquer outro tipo de sequência gráfica que não seja natural do português brasileiro apresentará transcrição, segmentação e marcação de tônica inadequadas no LexPorBR Infantil - Oral (Legendas).

5. Plataforma Web LexPorBR Infantil: construção e pesquisa

Com o objetivo de disponibilizar o máximo de informação sobre as palavras do LexPorBR Infantil - Oral (Legendas), 48 colunas com dados lexicais e metalinguísticas foram criadas e derivadas. As colunas com categorias de informações do LexPorBR Infantil - Oral (Legendas) são listadas na Tabela 3.

1. Lexema	13. Orto_freq/M	25. Lema_freq_log10	37. CVCV_sílaba
2. Fonologia	14. Orto_freq_log10	26. Lema_escala_zipf	38. Tipo_1
3. POS	15. Orto_escala_zipf	27. Nb_homógrafas	39. Tipo_2
4. POS_MM	16. Orto_zipf_rank	28. Vizinhos_ortográficos	40. Tipo_3
5. POS_DELAF	17. Fono_freq	29. OLD20	41. UNITEX
6. Sílabas/Tônica	18. Fono_freq_laplace	30. PUO	42. Hunspell
7. Nb_letras	19. Fono_freq_lexema/M	31. Nb_homófonas	43. Dicionário
8. Nb_fonemas	20. Fono_freq_log10	32. Vizinhos_fonológicos	44. Invertida_lexema
9. Nb_sílabas	21. Fono_escala_zipf	33. PLD20	45. Invertida_fono
10. Lema	22. Fono_zipf_rank	34. PUF	46. Invertida_lema
11. Orto_freq	23. Lema_freq	35. CVCV_lexema	47. Invertida_sílabas
12. Orto_freq_laplace	24. Lema_freq/M	36. CVCV_fonologia	48. Invertida_CVCV

Tabela 3. Categorias de informação do LexPorBR Infantil - Oral (Legendas).

As células número de letras/fonemas/sílabas foram calculadas através de um contador de caracteres; as células número de homógrafas/homófonas foram calculadas a partir de um contador de lexemas/fonologia repetidos, respectivamente; a célula fono_freq_laplace foi calculada a partir da frequência do córpus + 1, para comparação com outros córpus [Brysbaert and Diependaele 2012]; a célula fono_freq/M, contendo a frequência da palavra entre 1 milhão de palavras, foi calculada a partir da divisão da frequência de La Place pelo total de *tokens* do córpus; a célula fono_freq_log10 foi calculada a partir do log base 10 das frequências por milhão, com o objetivo de se linearizar a distribuição das frequências; as células fono_escala_zipf e fono_zipf_rank foram calculadas com objetivo de comparação das frequências entre diferentes córpus usando uma distribuição linearizada e ranqueada, respectivamente [van Heuven et al. 2014]. As células correspondentes às frequências ortográficas foram derivadas do Léxico do Português Brasileiro [Estivalet and Meunier 2015]. As células vizinhos ortográficos/fonológicos foram calculadas através da comparação de cada entrada lexical com todas as demais palavras do córpus a partir da distância de Hamming = 1 (substituição de 1 letra por vez); as células distância ortográfica/fonológica de Leveinshtein 20 (OLD20/PLD20) apresentam uma medida mais flexível de semelhança lexical a partir do cálculo do número de inserções, exclusões ou substituições das 20 palavras mais próximas [Yarkoni et al. 2008]. Para uma maior precisão destas normas, estas quatro categorias foram calculadas primeiramente entre as palavras existentes em pelo menos um dicionário dentre os utilizados neste trabalho e posteriormente para as demais entradas lexicais do córpus. As células ponto de unicidade ortográfico/fonológico (PUO-PUF) apresentam a informação sobre a partir de que letra/fonema a entrada lexical é única no léxico através de uma comparação com a entrada anterior e posterior no córpus organizado em ordem alfabética. As células CVCV apresentam a estrutura de consoantes e

vogais das formas, calculadas substituindo-se as vogais por V e consoantes por C. As células invertidas apresentam as respectivas entradas lexicais na forma invertida. Enfim, a célula POS, contendo 9 etiquetas gerais (ADJ, ADV, FUNC, IN, N, NUM, PCP, PRO, V), foi criada através da simplificação das 25 etiquetas específicas da célula POS_MM; considerou-se PROSUB = N, PROADJ = ADJ, CUR = NUM, ART + CONJ + PDEN + PREP + PRO = FUNC e desconsideram-se as contrações, como por exemplo ADV-KS = ADV. A célula POS_DELAF, contendo informações sobre flexão nominal e flexão verbal foi derivada a partir das definições do dicionário UNITEX-PB DELAF; as células Tipo_1/Tipo_2/Tipo_3/UNITEX/Hunspell apresentam uma marcação binária marcando se a entrada lexical está presente nestes materiais; ainda, a célula dicionário apresenta esta marcação se a palavra está presente em pelo menos um dentre estes cinco dicionários.

A interface apresenta dois motores de pesquisa: a pesquisa simples permite a inserção de uma lista de palavras a serem procuradas e a pesquisa complexa permite a inserção de uma série de especificações lexicais a serem procuradas ou evitadas.

6. Conclusões e Trabalhos Futuros

Os dois objetivos principais do presente estudo foram atingidos com êxito: i) criamos e disponibilizamos publicamente o LexPorBR Infantil - Oral (Legendas) com 129.053.297 *tokens* e 874.887 *types*; e ii) calculamos e derivamos 48 categorias de informações lexicais e metalinguísticas, contribuindo para o desenvolvimento de recursos lexicais carentes na pesquisa em psicolinguística e análise de córpus. O córpus criado apresenta alta riqueza lexical de um vocabulário oral de fala cotidiana derivado de legendas de filmes e séries familiares/infantis. No melhor do nosso conhecimento, este é o primeiro córpus baseado em palavras que apresenta as formas fonológicas e silábicas do português brasileiro. O LexPorBR Infantil - Oral (Legendas) pode ser acessado por interface web[4].

O LexPorBr Infantil pode ser usado de diferentes maneiras; seguem alguns exemplos abaixo.A base lexical é uma fonte valiosa de estímulos para estudos de aprendizagem e cognição de crianças. Portanto, pesquisadores da psicologia cognitiva, linguistas, neurocientistas, professores e outros podem se beneficiar da nossa base lexical para selecionar e combinar palavras de seu interesse e para monitorar suas propriedades psicolinguísticas. Isso tem sido feito em diferentes estudos sobre leitura [Schuster et al. 2015], memória de trabalho [Dominic et al. 2018], desenvolvimento de linguagem [Tomasello 2003] e muitos outros. Outra forma de utilização da base lexical é a exploração do próprio conteúdo para análise de redes semânticas, comparações com outras linguagens e geração de trajetórias interlinguísticas.

Os trabalhos futuros incluem a compilação de mais dois córpus para o LexPorBr Infantil: produzido por crianças e escrito (textos), para que tenhamos mais abrangência nos materiais, gêneros e frequências das palavras escutadas, lidas e escritas por crianças. Para estes córpus, pretende-se calcular a diversidade contextual. Também, pretendemos aprimorar o módulo de silabação e transcrição fonológica, enriquecer as entradas com informação morfológica e possibilitar a pesquisa dos contextos linguísticos originais (córpus de texto) onde as palavras ocorrem.

[4] `http://lexicodoportugues.com/infantil`

Referências

Aluísio, S., Pelizzoni, J., Marchi, A. R., de Oliveira, L., Manenti, R., and Marquiafável, V. (2003). An account of the challenge of tagging a reference corpus for brazilian portuguese. In *International Workshop on Computational Processing of the Portuguese Language*, pages 110–117. Springer.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., and Böhl, A. (2011a). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in german. *Experimental Psychology*, 58:412–424.

Brysbaert, M. and Diependaele, K. (2012). Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based choice. *Behavior Research Methods*, 45(2):422–430.

Brysbaert, M., Keuleers, E., and New, B. (2011b). Assessing the usefulness of google books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology*, 2:27.

Brysbaert, M. and New, B. (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Corral, S., Ferrero, M., and Goikoetxea, E. (2009). Lexin: a lexical database from spanish kindergarten and first-grade readers. *Behavior Research Methods*, 41(4):1009–17.

Dominic, G., Jean, S.-A., Gerald, T., and Anne, T. (2018). Does neighborhood size really cause the word length effect? *Memory cognition*, 46(2):244–260.

dos Santos, L. B., Duran, M. S., Hartmann, N. S., Candido, A., Paetzold, G. H., and Aluisio, S. M. (2017). A lightweight regression method to infer psycholinguistic properties for brazilian portuguese. In *International Conference on Text, Speech, and Dialogue*, pages 281–289. Springer.

Estivalet, G. L. and Meunier, F. (2015). The brazilian portuguese lexicon: An instrument for psycholinguistic research. *PLOS ONE*, 10(12).

Fonseca, E. R., Rosa, J. L. G., and Aluísio, S. M. (2015). Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society*, 21(1):2.

Hartmann, N. S., Paetzold, G. H., and Aluísio, S. M. (2018). Simplex-pb: A lexical simplification database and benchmark for portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 272–283. Springer.

Lambert, E. and Chesnet, D. (2001). Novlex: Une base de données lexicales pour les elèves de primaire [novlex: A lexical database for primary school children]. *L'Année Psychologique*, 2:215–235.

Lété, B., Sprenger-Charolles, L., and Colé, P. (2004). Manulex: A grade-level lexical database from french elementary school readers. *Behavior Research Methods*, 36:156–66.

Muniz, M. C. M. (2004). A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB. Master's thesis, ICMC-USP, São Carlos.

Paetzold, G. and Specia, L. (2016). Collecting and exploring everyday language for predicting psycholinguistic properties of words. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1669–1679, Osaka, Japan. The COLING 2016 Organizing Committee.

Schuster, S., Hawelka, S., Richlan, F., Ludersdorfer, P., and Hutzler, F. (2015). Eyes on words: A fixation-related fmri study of the left occipito-temporal cortex during self-paced silent reading of words and pseudowords. *Scientific Reports*, 5(12686).

Serrani, V. M. (2015). *Ambiente web de suporte à transcrição fonética automática de lemas em verbetes de dicionários do português do Brasil*. PhD thesis.

Silva, D. C., de Lima, A. A., Maia, R., Braga, D., de Moraes, J. F., de Moraes, J. A., and Resende, F. G. (2006). A rule-based grapheme-phone converter and stress determination for brazilian portuguese natural language processing. In *2006 International Telecommunications Symposium*, pages 550–554. IEEE.

Silva, D. d. C. (2011). Algoritmos de processamento da linguagem e síntese de voz com emoções aplicados a um conversor texto-fala baseado em hmm. *Doutorado, Programa de Engenharia Elétrica, Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia (COPPE/UFRJ), Rio de Janeiro*.

Soares, A. P., Medeiros, J. C., Simões, A., Machado, J., Costa, A., Iriarte, Á., de Almeida, J. J., Pinheiro, A. P., and Comesaña, M. (2014a). ESCOLEX: A grade-level lexical database from european portuguese elementary to middle school textbooks. *Behavior Research Methods*, 46(1):240–253.

Soares, A. P., Medeiros, J. C., Simões, A., Machado, J., Costa, A., Iriarte, Á., de Almeida, J. J., Pinheiro, A. P., and Comesaña, M. (2014b). Escolex: A grade-level lexical database from european portuguese elementary to middle school textbooks. *Behavior Research Methods*, 46(1):240–253.

Tang, K. (2012). A 61 million word corpus of Brazilian Portuguese film subtitles as a resource for linguistic research. *UCL Working Papers in Linguistics*, 24:208–214.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press, Cambridge, MA, US.

van Heuven, W. J. B., Mandera, P., Keuleers, E., and Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, 67(6):1176–1190.

Yarkoni, T., Balota, D., and Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15(5):971–979.