doi:http://dx.doi.org/10.5007/1984-8412.2016v13n1p1068

A DIVERSIDADE LEXICAL NA ESCRITA DE TEXTOS ESCOLARES

LA DIVERSIDAD LÉXICA EN LA ESCRITA DE TEXTOS ESCOLARES

LEXICAL DIVERSITY IN SCHOOL WRITTEN TEXTS

Mário Martins *
Universidade Federal do Amapá

RESUMO: Este artigo apresenta um estudo de correlação entre a diversidade lexical e a progressão escolar em textos escritos por crianças e adolescentes em idade escolar monolíngues de português europeu. Para mensurar a diversidade, utiliza-se a medida D (RICHARDS; MALVERN, 1997), versão matematicamente corrigida da medida TTR (TEMPLIN, 1957), que consiste na razão entre palavras diferentes (*types*) e palavras totais (*tokens*). Com o recurso às ferramentas CLAN (MACWHINNEY, 2000) e IMS Open Corpus Workbench (EVERT; HARDIE, 2011), esta medida foi aplicada a um *corpus quasi*-longitudinal, com 244 textos de registros narrativos (n=122) e argumentativos (n=122), escritos por alunos do quinto (n=26), do sétimo (n=46) e do décimo (n=50) anos do sistema escolar português. Os resultados mostram que, em ambos os registros, há uma correlação positiva entre a progressão escolar e o desenvolvimento lexical, mas que não se mostra linear de um ano a outro, particularmente do quinto ao sétimo ano. Para fundamentar a discussão, são apresentados dois exemplos dos modos como os textos escolares sob estudo variam quanto à utilização do vocabulário. Pretende-se, com este trabalho, contribuir para uma compreensão mais pormenorizada dos movimentos configuradores do desenvolvimento da língua escrita de crianças e jovens em idade escolar.
PALAVRAS-CHAVE: Diversidade lexical. Progressão escolar. Textos escolares.

RESUMEN: Este artículo presenta un estudio de correlación entre la diversidad léxica y la progresión en la escuela en textos escritos por niños y adolescentes en edad escolar monolingües de portugués europeo. Para caracterizar la diversidad, se utiliza D (RICHARDS; MALVERN, 1997) como medida, una versión corregida matemáticamente de la medida TTR (TEMPLIN, 1957), que significa la razón entre las diferentes palabras (*types*) y las palabras totales (*tokens*). Con el uso de la herramienta CLAN (MACWHINNEY, 2000) y IMS Open Corpus Workbench (EVERT; HARDIE, 2011), esta medida se aplicó a un *corpus* cuasi longitudinal, con 244 textos de registros narrativos (n = 122) y argumentativos (n = 122), escritos por estudiantes de quinto (n = 26), séptimo (n = 46) y décimo (n = 50) grado del sistema escolar portugués. Los resultados muestran que en ambos registros, existe una correlación positiva entre la progresión en la escuela y la diversidad léxica, pero no se muestra linear de un año a otro, sobre todo a

* Doutor em Linguística Educacional pela Universidade de Lisboa (UL), professor da Universidade Federal do Amapá (UNIFAP) e coordenador do Grupo de Pesquisa sobre o Ensino de Línguas para Fins Específicos (GP LIFE). E-mail: mgcmartins@gmail.com.

partir del quinto al séptimo grado. Para apoyar el debate, se presentan dos ejemplos de las formas como los textos en el estudio varían en el uso del vocabulario. El objetivo de este estudio contribuye a una mejor comprensión de los trayectos del desarrollo del lenguaje escrito de los niños y jóvenes en edad escolar.
PALABRAS CLAVE: Diversidad léxica. Progresión escolar. Textos escolares.

ABSTRACT: This article presents a correlational study between lexical diversity and school progression in texts written by school age children and adolescents, monolingual speakers of European Portuguese. The measure used to assess lexical diversity is D (RICHARDS; MALVERN, 1997), a mathematical correction of the TTR measure (TEMPLIN, 1957), which is based on the ratio of different words (types) to the total number of words (tokens). Using CLAN (MACWHINNEY, 2000) and IMS Open Corpus Workbench (EVERT; HARDIE, 2011) tools, this measure was applied to a quasi-longitudinal *corpus* consisting of 244 texts of narrative (n = 122) and argumentative (n = 122) register, written by students in the fifth (n = 26), the seventh (n = 46) and tenth (n = 50) year of the Portuguese basic schooling system. The results show that there are positive correlations between lexical diversity and school progression in both registers, although not linear from the fifth to the tenth year. To support the discussion, two examples of vocabulary variation are presented. This work aims to contribute to a more detailed understanding of the lexical development of children and adolescents written language across school progression.
KEYWORDS: Lexical diversity. School progression. School texts.

1 INTRODUÇÃO

Como afirmam Wray e Medwell (2006, p. 17), estudos de base empírica sobre o desenvolvimento lexical nos anos finais da infância e na adolescência são surpreendentemente raros, se comparados com os estudos sobre a aquisição nos primeiros anos da infância. No entanto, Berman (2007, p. 348) lembra que é justamente este conhecimento o componente mais saliente do desenvolvimento linguístico de crianças e adolescentes em idade escolar. Neste período, não apenas um grande número de novas palavras é gradualmente incorporado no seu repertório, como também é adquirida a capacidade de expressar uma grande variedade, suportada, em especial, no uso de palavras de maior extensão e de menor recorrência no discurso cotidiano, advindas, por vezes, de registros especializados.

Daller et al. (2007, p. 8) apresentam o conhecimento lexical de um aprendiz como um processo tridimensional, que compreende a extensão, que se refere à quantidade de palavras, a profundidade, que se refere à variedade do vocabulário, e a fluência, que se refere à velocidade de memorização. Em linha semelhante, Read (2000, p. 197-198) propõe que o desenvolvimento lexical[1] seja avaliado por quatro indicadores: a diversidade, que se mede pela razão entre a ocorrência de palavras diferentes e a totalidade de palavras num texto (também conhecida, no inglês, como *type-token ratio*); a densidade, que é a razão entre o número de palavras lexicais e o de palavras totais; a raridade, que se calcula pela razão entre a quantidade de palavras de frequência incomum e a de palavras totais; e o número de erros, que se relaciona com falhas na ortografia, na flexão ou na derivação ou com a interferência de outras línguas.

Para a avaliação do desenvolvimento lexical, terei em consideração um dos indicadores referidos por Read (2000): a diversidade, que é, pela sua natureza quantitativa, bastante adequada a um estudo baseado em *corpus*, além de já estar estabilizada na literatura enquanto indicador fiável de verificação do desenvolvimento linguístico do repertório lexical. A diversidade lexical consiste na razão entre a frequência de palavras diferentes e a frequência de itens totais (TEMPLIN, 1957), na sua versão matematicamente corrigida, conhecida como medida *D* (RICHARDS; MALVERN, 1997).

Estabeleço como principal hipótese de trabalho que a diversidade nos textos dos registros narrativo e argumentativo aumenta conforme o aluno avança nos anos escolares. Esta hipótese assenta na ideia comum (e sacralizada) de que, na escola, os processos de ensino e aprendizagem da escrita, e consequentemente do léxico, são graduais e ascendentes. Esta ideia reproduz-se no Programa

[1] Read (2000, p. 200 e seguintes) não usa a expressão "desenvolvimento lexical", mas "riqueza lexical" (no original, *lexical richness*). Por acreditar que a expressão "riqueza lexical" pode ser interpretada como pertencente a um domínio investigativo que ultrapassa os domínios deste texto, opto por usar a expressão "desenvolvimento lexical", mesmo que ainda não esteja devidamente estabilizada na literatura sobre o desenvolvimento linguístico em idade escolar. Pela mesma razão, opto por "raridade lexical", em linha com Malvern et al. (2004, p. 5), em vez de "sofisticação lexical" (no original, *lexical sophistication*).

de Português do Ensino Básico, em que se afirma que: "[...] o processo de ensino e aprendizagem do idioma progride por patamares sucessivamente consolidados. De acordo com esta noção, a aprendizagem constitui um 'movimento' apoiado em aprendizagens anteriores." (PORTUGAL, 2009, p. 9-10)[2]

Naturalmente, não se deduz de tal afirmação que o desenvolvimento seja sistemático e homogêneo para cada um dos alunos, mas sim que é possível distribuir este desenvolvimento em estágios discretos, mais ou menos ordenados, estando os alunos localizados em alguma parte do contínuo que vai de um estágio inicial a um estágio final. Como hipótese secundária, parto do princípio de que os textos do registro narrativo, em todos os anos escolares, apresentam maior diversidade que os textos do registro argumentativo. Justifica esta hipótese o fato de a inserção deste tipo de textos ocorrer desde os anos iniciais do primeiro ciclo da escolaridade, mantendo-se como uma necessidade de aprendizagem até, pelo menos, o 7.º ano, onde tais textos figuram para corroborar as "[...] práticas de relato e reconto de experiências, de acontecimentos, de filmes vistos ou de livros lidos" (PORTUGAL, 2009, p. 173), enquanto conteúdo formalmente proposto para o ensino das habilidades de escrita, aparecendo os textos argumentativos como uma preocupação do ensino da escrita apenas a partir do segundo ciclo da escolaridade.

2 DIVERSIDADE

A diversidade lexical é um indicador de desenvolvimento linguístico associado à quantificação da variação de palavras empregues num dado texto, ou seja, quanto maior a variação de palavras, maior a diversidade. Ransdell e Wengelin (2003 apud MCNAMARA et al., 2010, p. 57) sugerem que quanto maior é a diversidade lexical, mais patente é a competência linguística do falante/escritor. Por isso, pode-se concluir que um conhecimento limitado do vocabulário conduz à repetição e, consequentemente, reduz a complexidade de um texto. McCarthy e Jarvis (2010, p. 382) consideram que uma pequena taxa de diversidade lexical pode indicar a saturação temática num dado texto, ou seja, sem novas palavras, não há a introdução de novos temas.

Sobre o desenvolvimento linguístico em idade escolar, vários estudos analisam a diversidade lexical produzida por crianças e adolescentes. São referências, por exemplo, os estudos de Berman e Verhoeven (2002), Stromqvist et al. (2002) e Johansson (2009), originados do projeto Spencer (BERMAN; VERHOEVEN, 2002), cujo objetivo era examinar, em sete línguas diferentes (inglês, holandês, francês, islandês, hebreu, espanhol e sueco), as competências em língua materna não apenas de crianças e adolescentes, mas também de adultos, cobrindo uma faixa etária que se estende dos nove aos trinta anos. No contexto português, podem ser citados os trabalhos de Rodrigues (2008), que verifica a diversidade lexical num *corpus* de textos narrativos escritos por crianças monolíngues do 1.º ao 4.º ano, e Costa (2010), que analisa a diversidade lexical em produções textuais de alunos do 4.º, 6.º e 9.º anos.

Tão variadas quanto as aplicações da diversidade lexical são as fórmulas para a sua aferição. A medida clássica para medir este indicador encontra-se em Templin (1957), para quem a diversidade consiste na razão entre o número total de diferentes palavras, isto é, *types,* e o número total de palavras, *tokens*[3]. Esta medida é também conhecida simplesmente por TTR, forma abreviada da expressão em inglês *type-token ratio*[4]. Para que um texto apresente uma alta taxa de diversidade lexical, o escritor deve recorrer a palavras variadas e evitar a repetição, motivo pelo qual a diversidade lexical é habitualmente associada à noção de produtividade (WAGNER et al., 2011, p. 203). Os excertos abaixo exemplificam a aplicação da medida TTR (com *types* em itálico):

1. *Sim as redes sociais* sim *são importantes hoje em dia para a gente comunicar*. Sim *eu sou* a *favor porque* as redes sociais são *precisas tal como o Facebook e* a gente *fala por ele com* a *família* e com *os amigos*. E *primos* e *tias* e *etc...* e *muita mais* gente. (gpts_2_5c_mda)

[2] É importante referir que, no ano de 2015, foi homologado um novo programa de português (PORTUGAL, 2015), mas que conserva a ideia de progressão como se vê no programa de 2009.

[3] No programa Clan, considera-se o item *a* um *type* diferente do item *à* ou o item *rede*, por exemplo, diferente do item *redes*.

[4] Para o português, Berber-Sardinha (2004, p. 94) propõe alternativamente as seguintes traduções para *type-token: forma-item* ou *vocábulo-ocorrência*. Por acreditar que nenhuma destas expressões já se tenha estabilizado plenamente na língua portuguesa, opto por utilizar as expressões em inglês.

2. *As redes sociais*, *hoje em dia*, *são um importante meio de comunicação*, *servindo também para o entretenimento das pessoas*. O *Windows Live Messenger permite-nos falar com amigos* em *tempo real*, *e* para *mim é uma* das *únicas maneiras que tenho* de falar com pessoas que *moram longe ou* que *já não vejo há muito* tempo. (ls2_2_10a_fdl)[5]

O excerto (1), extraído de um texto argumentativo do quinto ano, apresenta-se com 35 *types* distribuídos em 52 *tokens*, o que resulta numa taxa TTR de 0,67. No excerto (2), extraído de um texto argumentativo do décimo ano, há um total de 44 *types* para 56 *tokens*, o que resulta numa taxa TTR de 0,79. Logo, o excerto argumentativo, a considerar as taxas obtidas, é mais diverso no emprego do vocabulário, sustentando-se menos na repetição vocabular.

Apesar da aparente eficiência da medida TTR em revelar a diversidade lexical de um texto, uma relação dependencial e progressiva entre esta medida como proposta por Templin (TTR = V/N, onde V corresponde a número de *types* e N a número de *tokens*) e a extensão do texto está reportada em McCarthy e Jarvis (2010, p. 381). Em termos práticos, à medida que um texto avança, mais palavras, ou itens, vão sendo incorporadas nele, implicando, portanto, maiores quantidades de *tokens*, mas a taxa de crescimento de palavras diferentes diminui proporcionalmente, ou seja, quanto maior for um texto, menor será a TTR. Dito de outro modo, quanto mais se avança num texto, menos se pode extrair dele em relação ao seu vocabulário, o que se explica pela necessidade óbvia que um escritor tem de ser coesivo, de ser hábil em equilibrar o fluxo de informações dadas e informações novas, razão por que não pode abdicar totalmente dos dispositivos de coesão, entre os quais se encontra a repetição de vocábulos.

O problema imposto pela técnica TTR tem implicações na metodologia de uma investigação. Conscientes do problema, alguns investigadores, como Biber (1995), por exemplo, limitam a análise a porções reduzidas dos textos do *corpus*, ou limitam-na a um número específico de *tokens*. Em ambos os casos, não se consideram os textos na sua totalidade. Além disso, há autores, como Ertmer et al. (2002 apud MCCARTHY; JARVIS, 2010, p. 381), que nem referem a problemática da extensão do texto, o que coloca em causa os resultados demonstrados.

Para contornar os inconvenientes impostos pela medida TTR, outras formas de cálculo são propostas, como faz, por exemplo, Guiraud (1960), que corrige a relação dependencial entre a TTR e a extensão do texto, quer pelo recurso à raiz quadrada, quer a algoritmos. No entanto, Durán et al. (2004, p. 221) garantem que essas reformulações matemáticas não superam o problema. Mais recentemente, Richards e Malvern (1997) propõem a medida *D*, que prevê a potencial redução da diversidade lexical em textos mais longos e, consequentemente, permite a comparação de textos de extensões distintas. A medida baseia-se num modelo de probabilidade para rastrear a diminuição da TTR conforme o número de *tokens* aumenta. Quanto maior for o valor *D*, maior será a diversidade lexical[6]. Se aplicada aos mesmos excertos acima, (1) e (2), têm-se, respectivamente, as seguintes taxas: 36,49 e 96,08.

Como os textos escolares do *corpus* aqui investigado se apresentam com extensões variadas, opto pela medida *D*. São referências na aplicação desta medida como indicador do desenvolvimento lexical os estudos de Berman e Verhoeven (2002), Stromqvist et al. (2002) e Johansson (2009). Para o português europeu, não encontrei estudos focados na mesma população que utilizassem a medida *D*.

3 MÉTODOS E *CORPUS*

A fim de avaliar a correlação entre a diversidade e a progressão nos anos escolares, instituiu-se como referencial cada um dos três ciclos da escolaridade básica do sistema educacional português. Deste modo, selecionaram-se alunos regularmente matriculados no 5.º ano, a representar o desenvolvimento lexical ocorrido no 1.º ciclo; alunos matriculados no 7.º ano, a representar o desenvolvimento lexical no 2.º ciclo; e alunos matriculados no 10.º ano, a representar o desenvolvimento lexical ocorrido no 3.º

[5] Todos os arquivos do *corpus* foram nomeados com a seguinte composição: as letras iniciais do nome do aluno (p. ex.: abrm), o registro (1 - narração ou 2 - argumentação), o ano escolar e a turma (p. ex.: 5c) e a escola de origem (p. ex.: mda).

[6] Diferentemente da medida TTR, em que 1 é o máximo, a medida *D* não tem um limite preestabelecido.

ciclo. Todos os alunos são de escolas públicas localizadas no distrito de Lisboa.

A cada participante, foram aplicados dois estímulos: um para a obtenção de um texto narrativo e outro para a obtenção de um texto argumentativo, estando em conformidade com Nippold (2004, p. 2), segundo a qual o desenvolvimento linguístico se revela melhor em textos completos, pelo que os textos narrativos, expositivos ou argumentativos são os mais adequados para a manifestação das marcas do desenvolvimento linguístico, tendo em conta que os falantes/escritores, nesses textos, entram em ações comunicativas plenas, o que lhes exige um esforço linguístico (e cognitivo) bastante mais substancial do que em estímulos constituídos por pares de perguntas e respostas, por exemplo. Excluídos os textos de alunos não monolíngues de português europeu, os textos restantes foram armazenados informaticamente. Em virtude dos fins deste estudo e da necessidade de que houvesse um adequado reconhecimento pela plataforma Clan, foram corrigidos desvios de natureza gráfica, tais como acentuação, capitalização ou indentação. As informações gerais sobre os participantes e sobre o *corpus*, intitulado CODES (MARTINS, 2015), descrevem-se na tabela abaixo:

	5.°	**7.°**	**10.°**	**Total**
Idade (M, DP)	10,19 (0,402)	12,33 (0,701)	15,16 (0,370)	—
Sexo **feminino (%)** **masculino (%)**	 18 (69,2%) 8 (30,8%)	 25 (54,3%) 21 (45,7%)	 29 (58%) 21 (42%)	 72 (59%) 50 (41%)
Média em Português (M, DP) **no ano anterior (escala 0-5)**	3,92 (0,392)	3,46 (0,808)	3,84 (0,681)	—
N.º total de palavras	8.586	15.239	19.499	4.324
N.º total de textos	52	92	100	244

Tabela 1: Informações gerais sobre os participantes do estudo e sobre o *corpus* CODES.

Para a obtenção dos valores da diversidade lexical, os textos, convertidos para o formato CHA, foram inseridos na plataforma *Clan* (MACWHINNEY, 2000), de que se obtiveram os valores por texto da medida *D* (programa vocd). Para a extração de uma lista de *tokens* por ano e por registro textual, utilizou-se o programa *IMS Open Corpus Workbench* (EVERT; HARDIE, 2011), que, baseado no processador *CQP* (*Centralized Query Processor*), consiste numa coleção de ferramentas de código aberto que se destina a fazer consultas em *corpora* de grande extensão. Os textos do *corpus* foram previamente anotados morfossintaticamente com o MBT (*memory-based tagger*) (DAELEMANS et al., 1996), treinado para o português sobre uma parte do *corpus* CINTIL. Para a lematização, submeteu-se o *corpus* à versão portuguesa do MBLEM (VAN DEN BOSCH; DAELEMANS, 1999). Os dados, delimitados por tabulação, foram, de seguida, importados para o programa *Numbers*, onde foi possível configurar uma lista das ocorrências de *types* com as respectivas frequências.

Considera-se, para todos os efeitos estatísticos, o ano escolar (5.º, 7.º e 10.º) combinado com o registro textual (narrativo ou argumentativo) como variável independente, já que a seleção vocabular é sensível ao registro (JOHNSON; JOHNSON, 1999, p. 151). Como variável dependente, para avaliar a diversidade, consideraram-se os valores da medida *D.* Dois testes estatísticos, no programa *SPSS*, foram aplicados aos valores obtidos: a) testes de correlação (Pearson) e b) análise de variância (ANOVA), com *post-hoc* de Tukey.

Antes da aplicação desses testes estatísticos, os valores da medida *D* foram verificados quanto à sua distribuição, se normal ou não, com base no teste de Shapiro-Wilk e na inspeção visual dos histogramas resultantes. A normalidade dos dados por este teste mede-se por um valor de $p \geq 0,05$. Constatou-se que os valores constituintes da medida *D,* tanto no registro narrativo como no registro argumentativo, apresentavam uma distribuição anormal, sendo normalizados pela operação Log10. É sobre os dados transformados que se aplicam os testes estatísticos.

4 RESULTADOS

Na Tabela 2, abaixo, apresentam-se, divididos por registro (narrativo e argumentativo), os valores médios de ocorrência de itens totais (*tokens*) e de itens diferentes (*types*), com respectivos desvios-padrão, identificados nos textos sob estudo. São esses os constituintes fundamentais da medida *D*, cujos valores, também diferenciados por registro, se apresentam mais abaixo, no Gráfico 1, seguindo-se a apresentação e a discussão dos resultados dos testes estatísticos.

Registro	Ano	N	Tokens (DP)		Types	(DP)
Narrativo	5.º	26	179,12	45,14	101,62	19,53
	7.º	46	180,54	37,32	104,50	16,89
	10.º	50	207,62	48,40	129,20	25,25
Argumentativo	5.º	26	154,46	39,26	88,58	17,78
	7.º	46	153,24	38,26	91,61	20,80
	10.º	50	185,54	32,11	113,56	18,15

Tabela 2: Frequência média de itens totais (*tokens*) e itens diferentes (*types*) identificados nos registros narrativo e argumentativo.

Quanto à medida *D* identificada no registro narrativo, os valores médios crescem de ano a ano. No quinto ano, o valor médio obtido é de 78,27, com um intervalo de ocorrências que se limita, no mínimo, por 47,73 e, no máximo, por 117,10; no sétimo ano, o valor médio é de 83,35, com um intervalo de ocorrências que se limita, no mínimo, por 37,43 e, no máximo, por 206,24; no décimo ano, o valor médio é de 109,87, com um intervalo de ocorrência que se limita, no mínimo, por 65,23 e, no máximo, por 188,56. Quanto ao registro argumentativo, no quinto ano, o valor médio de D é 74,44, com um intervalo de ocorrências que se limita, no mínimo, por 44,57 e, no máximo, por 129,72; no sétimo ano, o valor médio é de 83,94, com um intervalo de ocorrências que se limita, no mínimo, por 38,75 e, no máximo, por 143,77; no décimo ano, o valor médio é de 100,76, com intervalo de ocorrência que se limita, no mínimo, por 66,86 e, no máximo, por 157,27.

No Gráfico 1, é possível ilustrar o movimento de elevação da medida *D* do quinto para o sétimo e deste para o décimo ano, em ambos os registros, sendo a mudança aparentemente mais expressiva apenas a partir do sétimo ano:

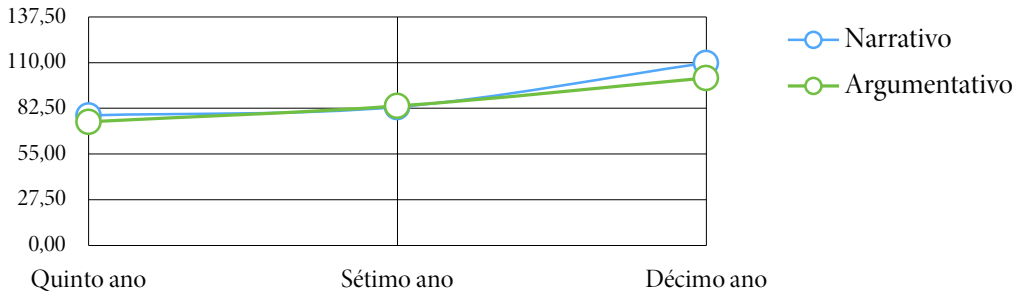

Gráfico 1: Distribuição dos valores médios da medida D, em números absolutos, nos registros narrativo e argumentativo ao longo dos anos escolares.

Os testes de correlação indicam que, nos textos narrativos, há uma correlação positiva moderada[7] (r=0,506; p=0,000[8]) entre a medida *D* e a progressão escolar. Na análise da variância, corrobora-se que há uma diferença significativa entre os grupos estudados (F(22,924); p=0,000[9]), mas o teste *post-hoc* de Tukey indica que somente há diferenças significativas entre o quinto e o décimo ano (p=0,000) e entre o sétimo e o décimo (p=0,000), não havendo diferença significativa entre o quinto e o sétimo ano (p=0,780).

Quanto aos textos de registro argumentativo, há uma correlação positiva também moderada (r=0,453; p=0,000) entre a medida *D* e a progressão escolar. Na análise da variância, observa-se uma diferença significativa entre os anos escolares (F(15,358); p=0,000). Em termos específicos, o teste *post-hoc* de Tukey revela que há diferenças significativas tanto entre o quinto e o décimo ano (p=0,000), como há também entre o sétimo e o décimo ano (p=0,001), mas não há diferenças significativas entre o quinto e o sétimo ano (p=0,141).

4.4 EXEMPLIFICAÇÃO DA DIVERSIDADE LEXICAL

Como asseguram Gillis e Ravid (2009, p. 229), as mudanças por que passa a língua de uma criança ou adolescente não se configuram como fenômenos linguísticos isolados, mas sim relacionados com complexas transformações de natureza cognitiva, social, afetiva, comportamental e, acrescento, biológicas, as quais direta ou indiretamente podem justificar as diferenças identificadas nos textos escritos escolares estudados. Justificações desses tipos, no entanto, vão muito além do escopo deste texto. Sem o mesmo rigor estatístico da descrição dos indicadores, mas fornecendo informações de valores percentuais, é possível realizar breves menções sobre alguns casos de uso que os participantes deste estudo fazem do léxico, mas não somente, já que há implicações, por vezes, de natureza sintática associadas. Estes casos sobre usos linguísticos particulares colocam-se como potenciais pistas para investigações futuras, e com recortes diferentes do que aqui se propõe, sobre os modos como os textos se diversificam ao longo do desenvolvimento escolar. Para a apresentação de tais casos, toma-se, como ponto de partida, listas de palavras mais frequentes por ano e por registro, já que permitem a identificação de diferenças numéricas das ocorrências que podem ser tomadas como mais relevantes. As menções a seguir apresentadas ora se baseiam em palavras que figuram explicitamente nas listas, ora estão em relação estreita com alguma palavra que nela conste.

Veja-se, primeiramente, na tabela 3, a lista de palavras mais frequentes (por *tokens*) nos textos de registro narrativo, que, em valores percentuais, correspondem, no quinto ano, a 5,99% do total de palavras; no sétimo ano, a 5,56%; e, no décimo ano, a 3,75%:

Ranking	**Quinto ano**		**Sétimo ano**		**Décimo ano**	
	Palavra	Frequência	Palavra	Frequência	Palavra	Frequência
1	fomos	36	fomos	64	dia	55
2	ir	35	dia	64	verão	55
3	dia	34	casa	64	ir	48
4	casa	32	amigo	53	casa	43
5	praia	30	praia	50	praia	37
6	verão	26	verão	47	amigo	36

[7] Segundo Dancey e Reidy (2004), a força da correlação avalia-se pelos seguintes valores de coeficiente (Pearson): coeficiente 1 = correlação perfeita; coeficiente entre 0,7 e 0,9 = correlação forte; coeficiente entre 0,4 e 0,6 = correlação moderada; coeficiente entre 0,1 e 0,3 = correlação fraca; e coeficiente 0 = correlação nula.

[8] Correlação significativa no nível 0,01 (duas extremidades).

[9] Diferença significativa no nível <0,05.

7	amiga	22	amiga	41	amiga	34
8	disse	21	ir	30	noite	31
9	férias	17	tinha*	26	tinha	31
10	ver	15	férias	20	amigos	27

Tabela 3: Lista das dez palavras lexicais mais frequentes (*tokens*) no registro narrativo ao longo da progressão escolar.

Segundo Berber-Sardinha e Shimazumi (2003, p. 18), listas de palavras podem ser indicativas do tema, ou temas, mais recorrentes num dado *corpus*. Assim, da lista acima, pode-se inferir claramente que a seleção lexical dos participantes reflete o tema da tarefa de escrita, que, em síntese, pede a narração de uma aventura no último verão, de que devem tomar parte o narrador e o melhor amigo. Pode-se inferir ainda da lista de palavras que tal aventura acontece com o narrador acompanhado por um ou mais amigos, num dia das férias de verão e numa casa na praia, elementos que, do ponto de vista discursivo, tipificam a construção do enquadramento introdutório de narrativas (*introductory setting elements*) (BERMAN, 2004, p. 275). É importante referir que as tarefas de escrita são entendidas aqui como o principal contexto de situação de textos escolares, ou, como sugiro, a tarefa é o contexto de situação local (em oposição ao contexto de situação global, não linguístico).

A lista de palavras acima, *grosso modo*, revela fortes semelhanças entre os três anos da escolaridade quanto ao léxico selecionado, em particular, no que respeita ao modo como os alunos se vinculam à tarefa. Há, no entanto, algumas ocorrências, porque podem diferenciar um ano de outro, que merecem atenção mais pormenorizada, a começar pela recorrência do verbo "ir" nos textos narrativos, que está representado, enquanto verbo lexical, nas listas dos três anos de escolaridade pelos *types* "fomos" e "ir", quer como verbo pleno, quer como auxiliar[10].

A utilização sintática que se faz do verbo "ir" nos textos narrativos expressa-se numa das seguintes funções: a) verbo pleno, seguido de complemento preposicionado (oblíquo): "No outro dia *fomos para outra cidade*", "Nós também *fomos à praia* muitas vezes"; b) verbo auxiliar temporal seguido de verbo no infinitivo: "Depois de comermos *fomos ver* o pôr-do-sol"; c) verbo auxiliar aspectual seguido de verbo no gerúndio ou no infinitivo (neste último caso, mediado pela preposição 'a'): "à medida que *íamos andando* a paisagem *ia melhorando* cada vez mais", "Elas no caminho *iam a contar* as férias fantásticas que tiveram". A distribuição de uso destas construções nos textos narrativos ilustra-se no Gráfico 2, a seguir:

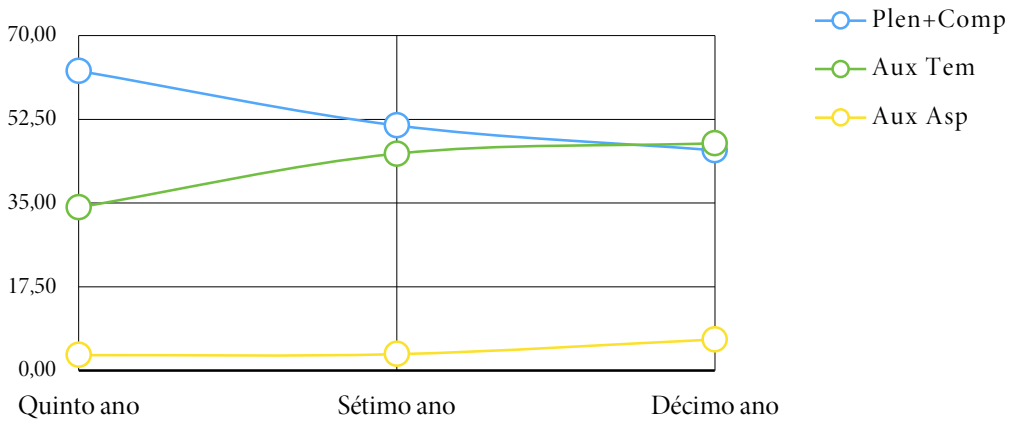

Gráfico 2: Distribuição dos valores médios de ocorrências, em percentual, de construções com o verbo 'ir' no registro narrativo ao longo dos anos escolares.

[10] Gonçalves e Costa concluem, com base num conjunto de testes sintático-semânticos, que o verbo "ir" é semi-auxiliar, dado que apresenta "[...] um comportamento duplo (por um lado idêntico ao dos auxiliares; por outro, idêntico ao de verbos principais, como parecer)" (2002, p. 80). Aqui, sigo Raposo (2013, p. 1221 e seguintes), que opta por incluí-lo na categoria dos verbos auxiliares.

Como se pode ver, a distribuição de usos não é plana ao longo da progressão escolar, havendo um expressivo declínio do emprego do verbo 'ir' em construções Plen+Comp, o qual, no quinto ano, representa 62,6% do total de ocorrências do verbo "ir"; no sétimo ano, representa 51,23% do total; e no décimo ano, 46,04%. Ao contrário disso, os auxiliares temporais e aspectuais crescem ano após ano. No quinto ano, 37,4% do total de ocorrências do verbo "ir" ocorrem com função auxiliar (temporal ou aspectual); no sétimo ano, este valor passa a 48,77%; e no décimo, a 53,96%.

A instanciação do verbo "ir" como auxiliar nos textos narrativos contribui para a perspectivação temporal dos eventos narrados, quer através da expressão de futuridade (como auxiliar temporal), servindo, de acordo com Cunha e Cintra (2010, p. 411), "[...] para exprimir o firme propósito de executar a ação, ou a certeza de que ela será realizada em futuro próximo", quer através da expressão da gradação do evento principal (como auxiliar aspectual), que serve, ainda segundo os gramáticos, "[...] para indicar que a ação se realiza progressivamente ou por etapas sucessivas" (2010, p. 411). Esta contribuição justifica-se porque, mesmo que este verbo seja o resultado de um processo de gramaticalização, ainda guarda a dimensão temporal do verbo pleno a que ele corresponde (RAPOSO, 2013, p. 1228).

A observação da lista de itens lexicais mais frequentes (*tokens*) nos textos argumentativos também permite um conjunto de menções sobre os usos diferenciados do décimo ano em relação aos outros anos. Veja-se a lista abaixo (Tabela 4):

Ranking	**Quinto ano**		**Sétimo ano**		**Décimo ano**	
	Palavra	Frequência	Palavra	Frequência	Palavra	Frequência
1	redes	60	redes	185	redes	222
2	sociais	56	sociais	171	sociais	208
3	*facebook*	51	pessoas	106	pessoas	139
4	pessoas	47	*facebook*	70	amigos	63
5	coisas	30	importantes	53	*facebook*	56
6	acho	29	amigos	51	Pessoa	34
7	amigos	27	dia	49	importantes	31
8	internet	27	acho	43	Vida	31
9	falar	24	coisas	35	Acho	30
10	fazer	24	falar	33	Falar	28

Tabela 4: Ranking das dez palavras lexicais mais frequentes no registro argumentativo ao longo da progressão escolar.

No registro argumentativo, as dez palavras mais frequentes do quinto ano correspondem a 9,69% do total de palavras (*tokens*); no sétimo ano, corresponde a 11,29%; e, no décimo ano, a 8,85%. E, semelhantemente ao que acontece no registro narrativo, a seleção

lexical nos textos argumentativos prende-se fortemente à tarefa de escrita, mas, diferentemente do caso anterior, a relação entre as palavras mais frequentes e a tarefa, que pede um texto de opinião sobre as redes sociais, não se realiza apenas por meio de potenciais associações semânticas, mas também por meio de repetição explícita de itens que constavam do enunciado da tarefa que lhes era solicitada, o que se exemplifica pela ocorrência dos nomes 'redes', 'sociais', '*facebook*', do adjetivo 'importantes' e do verbo 'achar', que constam da tarefa.

Um caso que merece ser mencionado refere-se ao emprego de alternativas ao verbo epistêmico 'achar', que tem ocorrência significativa nos três anos estudados. Verbos de valor modal epistêmico implicam "[...] graus de certeza ou avaliação de probabilidade acerca do conteúdo proposicional da frase" (OLIVEIRA; MENDES, 2013, p. 623). O verbo 'achar' enquadra-se no primeiro caso, já que expressa, nos textos argumentativos investigados, o grau de comprometimento do sujeito da frase (ou do escritor) em relação à importância da existência das redes sociais. Outros verbos que se podem considerar epistêmicos são 'julgar', 'crer', 'pensar', 'calcular', 'imaginar' e 'considerar'. Estes verbos foram identificados nos textos argumentativos, a fim de comparar a uma frequência com a frequência de 'achar', revelando potenciais indicadores do desenvolvimento lexical associado aos valores de modalidade expressos por meios verbais. O Gráfico 3 demonstra comparativamente a utilização destes verbos ao longo da progressão nos anos escolares:

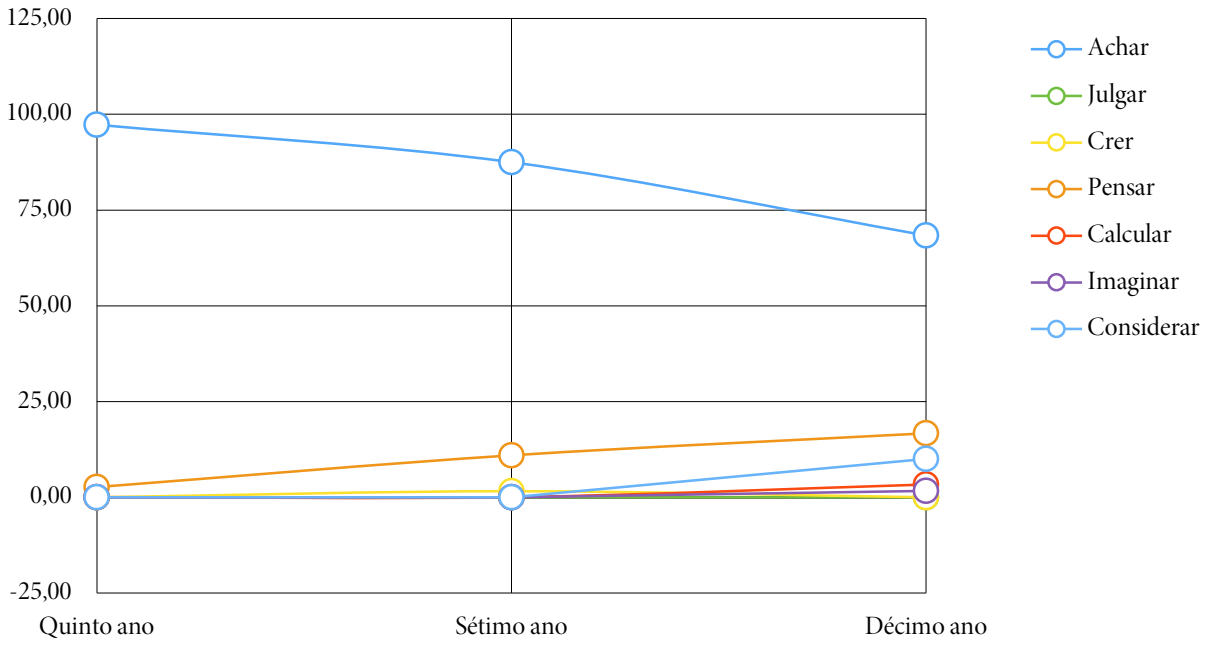

Gráfico 2: Distribuição dos valores médios de ocorrências, em percentual, de verbos modais epistêmicos no registro argumentativo ao longo dos anos escolares.

Nos textos argumentativos, como se pode inferir da lista de palavras mais frequentes e corroborar pelo Gráfico 3, a escolha prioritária para a expressão da crença, em todos os anos de escolaridade, é o emprego do verbo 'achar', que, no quinto ano, representa 97,30% das escolhas; no sétimo, 87,5%; e, no décimo ano, 68,33%. Estes valores percentuais decrescentes revelam também um acesso gradual a outras formas verbais possíveis para a expressão da modalidade epistêmica, de que se destaca o verbo 'pensar', que no quinto ano equivale a 2,7% do uso total, no sétimo ano, a 10,94% e no décimo ano, a 16,67%. Note-se ainda a ocorrência de 'considerar', apenas no décimo ano, em que representa 10% das ocorrências de verbos deste tipo.

5 DISCUSSÃO

Os resultados da aplicação dos testes de correlação apontam para uma tendência de crescimento da diversidade lexical, contabilizada por meio da medida *D*, ao longo da progressão do quinto para o décimo ano da escolaridade, tanto nos textos narrativos como nos textos argumentativos escritos pelos alunos participantes deste estudo, o que confirma a primeira hipótese de trabalho, segundo a qual a diversidade lexical aumenta conforme o aluno avança nos anos escolares. Todavia, como se viu pela aplicação dos testes *post-hoc*, o crescimento não é constante entre os três anos estudados, não se revelando diferenças significativas na diversidade lexical entre o quinto e o sétimo ano, nem no registro narrativo, nem no registro argumentativo. Quer isso dizer, em termos estritamente matemáticos, que a variação no uso do vocabulário identificada no quinto ano se assemelha à variação identificada no sétimo ano, havendo mudanças apenas na comparação com a variação no décimo ano da escolaridade, pelo que é possível afirmar que, em algum momento entre o sétimo e o décimo ano dos grupos estudados, se inicia, de fato, um processo de maior diversificação do uso das palavras.

Em Berman e Verhoeven (2002), Stromqvist et al. (2002) e Johansson (2009), reporta-se uma situação semelhante. Segundo esses estudos, há um significativo desenvolvimento da diversidade lexical, também examinada com base na medida *D*, apenas entre os treze e os dezessete anos de idade, sendo o desenvolvimento pouco significativo dos nove aos doze e dos dezessete aos trinta anos. Dois outros estudos, vistos em complementação, corroboram as conclusões dos estudos anteriores. Por um lado, Malvern et al. (2004), analisando textos produzidos por crianças e adolescentes em três diferentes grupos etários (7, 11 e 14 anos), notam que os resultados dos textos dos escritores mais velhos apresentam os resultados mais expressivos, sendo inexpressiva a diferença entre os dois grupos mais novos. Por outro lado, Crossley et al. (2011) identificam, em textos escritos também por três grupos etários (14-15 anos, 16-17 anos e primeiranistas universitários), diferenças estatísticas significativas entre todos os grupos. Para efeitos de comparação com os resultados aqui encontrados, vale a pena lembrar que, no percurso normal da escolaridade portuguesa, aos treze anos está-se no sétimo ano, início do segundo ciclo.

Para o português europeu, não foram identificados estudos que tratassem da diversidade lexical pela medida *D*, que aqui se considera, ou que fizessem a mesma abordagem quantitativa. Apesar disso, o estudo de Costa (2010) pode ser tomado como uma referência. Com o objetivo geral de examinar o uso de conectores de valor concessivo em diferentes tipos de produções escritas de alunos do quarto, sexto e nono anos da escolaridade básica, a investigadora (2010, p. 137-8) considera, enquanto estratégia complementar de caracterização quantitativa do *corpus*, a frequência média de *types* por texto, a partir do que identifica, do quarto para o nono ano, um aumento no uso de *types*. Costa observa que há um crescimento da diversidade lexical ao longo da progressão escolar, mas que, entre o quarto e o sexto ano, não se o vê especificamente.

Da síntese acima, é possível afirmar, em linha com Stromqvist et al. (2002, p. 53), que, entre os treze e os dezessete anos, há um salto expressivo na diversidade lexical da escrita escolar, pelo que se deve acautelar, como lembra Johansson (2009, p. 77), que as medidas de descrição da diversidade lexical entre grupos de pequenas diferenças de idade não devem ser usadas sem a complementação de outras medidas.

Ainda quanto aos resultados dos testes estatísticos aplicados ao *corpus* considerado, viu-se que a força da correlação entre a medida *D* e cada registro textual isoladamente confirma a segunda hipótese de trabalho, segundo a qual os textos do registro narrativo apresentam maior diversidade lexical que os textos do registro argumentativo. Como se viu, a força da correlação nos textos narrativos é mais expressiva do que nos textos argumentativos. Conforme Ravid (2004, p. 339), muitos estudos afirmam que, desde os cinco ou seis anos de idade (antes do período escolar, portanto), as crianças já possuem conhecimentos esquemáticos e linguísticos bem estabelecidos sobre a produção de textos narrativos, não se podendo afirmar o mesmo sobre os textos argumentativos, que só serão plenamente dominados na adolescência. Com isso, não se quer dizer que crianças e adolescentes não saibam distinguir um registro do outro, mas sim que lhes falta, até à adolescência, a familiaridade com o conjunto de conhecimentos linguísticos e retóricos que caracterizam o registro argumentativo, cujo foco é a apresentação de ideias e conceitos com base em argumentos e contra-argumentos (BERMAN; SLOBIN, 1994 apud RAVID, 2004, p. 339).

Os resultados de Berman e Ravid (2009), em complementação com o que aqui se identificou, parecem desafiar, pelo menos quanto à seleção lexical, a concepção, lembrada por Ravid (2004, p. 351), de que os textos narrativos, por serem adquiridos ainda nos anos iniciais da escolaridade e por serem aparentemente mais bem escritos, são mais complexos linguisticamente do que textos de outros registros. Isto não quer dizer que os textos argumentativos escritos pelos alunos portugueses sejam melhores ou mais bem escritos do que as suas contrapartes narrativas. Pode, ao contrário, apontar para o fato de estes alunos, quando confrontados com tarefas que lhes exigem uma ação verbal distinta daquela a que estão habituados, se esforçarem mais, demonstrando consciência sobre os mecanismos de produção dos registros textuais e das suas especificidades lexicais.

6 CONSIDERAÇÕES FINAIS

O estudo aqui reportado teve por objetivo central examinar a correlação entre um indicador de desenvolvimento da complexidade lexical, nomeadamente a diversidade, e a progressão nos anos escolares em dois registros textuais (narração e argumentação). É, portanto, um trabalho de natureza descritiva, pelo que tenta cumprir o seu objetivo ao apontar para valores estatísticos que podem ser, com as devidas ressalvas, tomados como tendências do desenvolvimento linguístico ao longo da evolução na escola.

Como se constatou, há uma correlação positiva entre a diversidade lexical (proporção de palavras *types* por *tokens*) e a progressão escolar do quinto ao décimo ano da escolaridade básica em ambos os registros textuais, ou seja, quanto mais se avança nos anos de escolaridade, ciclo a ciclo, mais variado se torna o vocabulário dos alunos. Esta correlação expressa-se mais fortemente no registro narrativo. No entanto, a correlação em ambos os registros caracteriza-se por não ser linear, já que não foram detectadas diferenças estatísticas significativas entre o quinto e o sétimo ano na quase totalidade dos testes.

O fato de não haver diferenças significativas no desenvolvimento da diversidade lexical em textos escritos entre o quinto e o sétimo ano, havendo apenas em relação ao décimo ano, pode ser surpreendente quando confrontado com a expectativa comum de que, na escola, de ano para ano, há um crescimento progressivo das competências vocabulares dos alunos. Na verdade, trata-se do reflexo de um persistente problema relacionado com o desenvolvimento da escrita escolar, bem sintetizado na seguinte afirmação de Wilkinson et al. (1980, p. 2): "*Development obviously takes place, but does not take place obviously*". Quer isto dizer que, apesar da certeza de que a escrita de um rapaz de 16 anos é provavelmente melhor do que a escrita de um rapaz de 11, ainda não se pode afirmar, com precisão, como esta melhoria, em termos desenvolvimentais, acontece.

Na tentativa de apontar para os modos como o desenvolvimento ocorre, apresentaram-se, adicionalmente ao estudo de correlação, dois casos relativos ao uso do léxico que pudessem ser tomados como pistas mais específicas do desenvolvimento lexical. Da observação destes casos, concluiu-se primeiramente que a tarefa escolar tem forte impacto na seleção lexical realizada pelos alunos, quer seja pela ativação de vocábulos semanticamente relacionados, quer seja pela repetição explícita de palavras. Concluiu-se ainda que a expressão do tempo, nos textos narrativos, é um indicador potencial de mudanças, em particular pela utilização de verbos auxiliares de natureza aspectual ou temporais. Nos textos argumentativos, como demonstrado, é relevante, no processo de desenvolvimento, a expressão da modalidade epistêmica pela seleção lexical que a realiza.

Este estudo aponta ainda para a necessidade de que, em estudos futuros, refinamentos deste indicador, como a diversidade baseada em lemas, ou outros indicadores de desenvolvimento da complexidade lexical, como a densidade, a raridade ou a extensão das palavras, por exemplo, sejam aplicados a *corpora* de desenho semelhante e maiores, a fim de se atestar se a ausência de diferenças entre o quinto e o sétimo ano se mantém ou não e, se não se mantém, qual é a natureza qualitativa da mudança.

Apesar de ser quase incontestável, como afirmam Gillis e Ravid (2009, p. 203), que uma criança que cresce num ambiente monolíngue tem acesso à maioria das estruturas morfológicas e sintáticas da sua língua antes de entrar na escola, há evidências acumuladas de que a aquisição da língua é um processo prolongado, que não se encerra aos dez ou doze anos, e de que consideráveis mudanças em todos os domínios linguísticos acontecem na língua de crianças mais velhas e adolescentes (BERMAN; VERHOEVEN, 2002). Isto, de fato, torna a língua dos adultos diferente da língua dos adolescentes, e a dos adolescentes diferente da língua de uma criança de doze anos. É, portanto, ao longo desse período da infância tardia e da adolescência, que grande parte

das características da língua adulta emerge e se consolida, acompanhada de construções de natureza mais complexa, cujas funções textuais são próprias de determinados tipos de textos a que só se tem acesso por meio da educação formal.

REFERÊNCIAS

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

BERBER-SARDINHA, T.; SHIMAZUMI, M. Schoolchildren writing: A corpus-based analysis. *Linguagem e Ensino*, v. 6, n. 1, p. 11-33, 2003.

BERMAN, R. A. The role of context in developing narrative abilities. In: STROMQVIST, S.; VERHOEVEN, L. (Ed.). *Relating events in narrative*: typological and contextual perspectives, Nova Jersey/Londres: Laurence Erlbaum, 2004. p. 261-280.

_____. Developing linguistic knowledge and language use across adolescence. In: HOFF, E.; SHATZ, M. (Ed.). *Blackwell handbook of language development*. Malden/Oxford/Victoria: Blackwell, 2007. p. 347-367.

BERMAN, R. A.; VERHOEVEN, L. Cross-linguistic perspectives on the development of text-production abilities: speech and writing. *Written Language and Literacy*, v. 5, n. 1, p. 1-44, 2002.

BERMAN, R. A.; RAVID, D. Becoming a literate language user: Oral and written text construction across adolescence. In: OLSON, D. O.; TORRANCE, N. (Ed.). *Cambridge handbook of literacy*. Cambridge: Cambridge University Press, 2009. p. 92-111.

BIBER, D. *Dimensions of register variation*: a cross linguistic comparison. Cambridge: Cambridge University Press, 1995.

COSTA, A. L. *Estruturas contrastivas*: desenvolvimento do conhecimento explícito e da competência de escrita. 2010. 306 f. Tese (Doutorado em Linguística Educacional) - Universidade de Lisboa, Lisboa, 2010.

CROSSLEY, S. A. et al. The development of writing proficiency as a function of grade level: a linguistic analysis. *Written Communication*, v. 28, n. 3, p. 282-311, 2011.

CUNHA, C.; CINTRA, L. *Nova gramática do português contemporâneo*. Rio de Janeiro: Lexikon, 2010.

DAELEMANS, W. et al. MBT: A memory-based part of speech tagger generator. In: Fourth Workshop on Very Large Corpora. Anais… Copenhagen: 1996, p. 14-27.

DALLER, H.; MILTON, J.; TREFFERS-DALLER, J. *Modelling and assessing vocabulary knowledge*. Cambridge: Cambridge University Press, 2007.

DANCEY, C.; REIDY, J. *Statistics without maths for psychology*: using SPSS for Windows. Londres: Prentice Hall, 2004.

DURÁN, P. et al. Developmental trends in lexical diversity. *Applied Linguistics*, v. 25, n. 2, p. 220-242, 2004.

EVERT, S.; HARDIE, A. Twenty-first century corpus workbench: updating a query architecture for the new millennium. In: PROCEEDINGS OF THE CORPUS LINGUISTICS, 2011, Birmingham. *Anais*… Birmingham: Universidade de Birmingham, 2011.

GILLIS, S.; RAVID, D. Language acquisition. In: DOMINIEK, S.; ÖSTMAN, J.; VERSCHUEREN, J. (Ed.) *Cognition and pragmatics*. Amsterdam/Filadélfia: John Benjamins, 2009. p. 201-249.

GONÇALVES, A.; COSTA, T. *(Auxiliar a) compreender os verbos auxiliares*. Lisboa: APP, Colibri, 2002.

GUIRAUD, P. *Problèmes et méthodes de la statistique linguistique*. Dordrecht: D. Reidel, 1960.

JOHANSSON, V. *Developmental aspects of text production in writing and speech*. Lund: Lund University, 2009.

JOHNSON, K.; JOHNSON, H. *Encyclopedic dictionary of applied linguistics*: a handbook for language teaching. Oxford/Malden: Blackwell, 1999.

MACWHINNEY, B. *The Childes Project*: tools for analyzing talk. Mahwah: Lawrence Erlbaum Associates, 2000.

MALVERN, D. et al. *Lexical diversity and language development: quantification and assessment*. Nova Iorque: Palgrave Macmillan, 2004.

MARTINS, M. Corpus Desenvolvimental: Codes. Lisboa: Centro de Linguística da Universidade de Lisboa (CLUL), 2015. Disponível em: <http://alfclul.clul.ul.pt/CQPweb/codes/>. Acesso em: 28 de janeiro de 2016.

MCCARTHY, P. M.; JARVIS, S. Mtld, vocd-d, and hd-d: a validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, v. 42, n. 2, p. 381-392, 2010.

MCNAMARA, D.; CROSSLEY, S.; MCCARTHY, P. Linguistic features of writing quality. *Written Communication*, v. 27, n. 1, p. 57-86, 2010.

NIPPOLD, M. A. Research on later language development. In: BERMAN, R. A. (Ed.) *Language development across childhood and adolescence*. Amsterdam/Filadélfia: John Benjamins, 2004. p. 1-8.

OLIVEIRA, F.; MENDES, A. Modalidade. In: RAPOSO, E. B. P. et al. (eds.) *Gramática do português*. Lisboa: Fundação Calouste Gulbenkian, 2013. p. 623-672.

PORTUGAL. *Programa de português do ensino básico*. Lisboa: Ministério da Educação e Ciência. Direção-Geral da Educação, 2009.

PORTUGAL. *Programa e metas curriculares de português do ensino básico*. Lisboa: Ministério da Educação e Ciência. Direção-Geral da Educação, 2015.

RAPOSO, E. B. P. Verbos auxiliares. In: RAPOSO, E. B. P. et al. (Ed.). *Gramática do português*. Lisboa: Fundação Calouste Gulbenkian, 2013. p. 1221-1284.

RAVID, D. Emergence of linguistic complexity in later language development: evidence from expository text construction. In: RAVID, D. D.; SHYLDKROT, H. B. (Ed.). *Perspectives on language and language development: essays in honor of Ruth A. Berman*. Dordrecht/Boston/Londres: Kluwer Academic Publishers, 2004.

READ, J. *Assessing vocabulary*. Cambridge: Cambridge University Press, 2000.

RICHARDS, B. J.; MALVERN, D. D. *Quantifying lexical diversity in the study of language development*. Reading: University of Reading, 1997.

RODRIGUES, S. B. P. *Escrita espontânea: desenvolvimento das capacidades de composição escrita em crianças do 1º ao 4º ano de escolaridade*. 2008. 133 f. Dissertação (Mestrado em Linguística Aplicada) - Universidade Fernando Pessoa, Porto, 2008.

STRÖMQVIST, S. et al. Toward a cross-linguistic comparison of lexical quanta in speech and in writing. *Written Language and Literacy*, v. 5, n. 1, p. 46-67, 2002.

TEMPLIN, M. *Certain language skills in children: their development and inter-relationships*. Minnesota: University of Minnesota Press, 1957.

VAN DEN BOSCH, A.; DAELEMANS, W. Memory-based morphological analysis. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 37, 1999, Maryland. *Anais*… MARYLAND: University of Maryland, 1999.

WAGNER, R. K. et al. Modeling the development of written language. *Reading and writing*, v. 24, n. 2, p. 203-220, 2011.

WILKINSON, A. et al. *Assessing language development*. Oxford: Oxford University Press, 1980.

WRAY, D.; MEDWELL, J. *Progression in writing and the northern ireland levels for writing*. Warwick: University of Warwick, 20

Recebido em 02/12/2015. Aceito em 23/01/2016.