DOI: 10.5433/2237-4876.2017v20n1p218

Densidade Lexical na Escrita de Textos Escolares

Lexical Density in School Written Texts

Mário **Martins***

Resumo: Este artigo apresenta um estudo de correlação entre a densidade lexical e a progressão escolar em textos escritos por crianças e adolescentes em idade escolar monolíngues de português europeu. Enquanto um indicador do desenvolvimento lexical, a densidade mensura-se pela razão entre as palavras lexicais e o total de palavras, como proposta por Ure (1971). Complementam-na a obtenção de razões baseadas em classes de palavras específicas, que são a densidade nominal, verbal, adjetival e adverbial. Com o recurso à ferramenta IMS Open Corpus Workbench, essas medidas foram extraídas de um *corpus quasi*-longitudinal, com 244 textos de registros narrativos (n=122) e argumentativos (n=122), escritos por alunos do 5º (n=26), do 7º (n=46) e do 10º (n=50) ano do sistema escolar português. Os resultados mostram que, em ambos os registros, não há correlação entre a progressão escolar e a densidade lexical se considerada indistintamente. Tomadas as medidas por classes de palavras, identificou-se, por um lado, uma correlação positiva entre a densidade nominal e a progressão escolar e, por outro, uma correlação negativa entre a densidade verbal e a progressão escolar. Pretende-se, com este trabalho, contribuir para uma compreensão mais pormenorizada dos percursos configuradores do desenvolvimento da língua escrita de crianças e jovens em idade escolar.
Palavras-chave: Densidade lexical. Progressão escolar. Textos escolares.

Abstract: This article presents a correlational study between lexical density and school progression in texts written by school age children and adolescents,

* Doutor em Linguística pela Universidade de Lisboa (UL). Professor da Universidade Federal Rural do Semi-Árido (UFERSA). Contato: mgcmartins@gmail.com.

monolingual speakers of European Portuguese. The measure used to assess lexical density is the ratio of lexical words to the total number of words (global density), as proposed by Ure (1971). This measure is complemented by the ratio of specific word classes (name, verb, adjective and adverb) to the total number of words. Using IMS Open Corpus Workbench tool, this measures was extracted from a *quasi*-longitudinal *corpus* consisting of 244 texts of narrative (n = 122) and argumentative (n = 122) register, written by students in the 5th (n = 26), the 7th (n = 46) and 10th (n = 50) year of the Portuguese basic schooling system. The results show that there are no correlations between global lexical density and school progression in both registers. Regarding specific lexical density, it was found that, in one hand, there is a positive correlation between nominal density and progression and, on the other hand, there is a negative correlation between verbal density and progression. 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 density. School progression. School texts.

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. Nesse 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, Milton e Treffers-Daller (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 produção e compreensão de palavras. Em linha semelhante, Read (2000, p. 197-198) sintetiza que o desenvolvimento lexical tipicamente é avaliado por quatro indicadores: a diversidade, a densidade, a raridade e o número de erros.

Para a avaliação do desenvolvimento lexical, terei em consideração um dos indicadores referidos por Read (2000): a densidade[1]. Proposto por Ure (1971), a densidade consiste na razão entre a frequência de palavras lexicais e a frequência de palavras totais, expressa por valores percentuais (STUBBS, 1996). Pela sua natureza quantitativa, este indicador é bastante adequado a um estudo baseado em *corpus*, além de já estar estabilizado na literatura enquanto indicador fiável de verificação do desenvolvimento linguístico do repertório lexical.

Estabeleço como principal hipótese de trabalho que a densidade nos textos de registro narrativo e argumentativo aumenta conforme o aluno avança nos anos escolares. Esta hipótese assenta na ideia comum (e sacralizada, para muitos) 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 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.

---

[1] Quanto à diversidade lexical em textos escolares, ver Martins (2016).

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

Como hipótese secundária, parto do princípio de que os textos do registro narrativo, em todos os anos escolares, apresentam maior densidade 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 7º 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.

1 Densidade

A densidade lexical é um dos principais indicadores da trajetória do desenvolvimento rumo a uma escrita mais acadêmica (COLOMBI, 2002, p. 72), relacionando-se com a capacidade que um escritor possui de dispor a informação e condensá-la, o que implica, portanto, nos modos como a informação representada pelo léxico é incorporada na estrutura gramatical. Desta maneira, um texto lexicalmente mais denso é um texto mais informativo, o que assenta no fato de serem os itens lexicais os portadores primários do conteúdo referencial (RAVID, 2004, p. 346).

A densidade se associa à capacidade cognitiva que crianças e jovens adquirem de usar a língua separada de contextos físicos imediatos, no que se convencionou chamar de língua descontextualizada, isto é, sem o suporte de contextos conversacionais (SNOW, 1983 apud SCHLEPPEGRELL, 2004, p. 8). Apesar das diversas críticas a que tal compreensão tem sido sujeita, a capacidade de produzir textos deslocados dos seus contextos reais, expressa, por exemplo, para referir o passado, antecipar o futuro ou especular sobre mundos possíveis alternativos à realidade, continua a ser bastante valorizada nos sistemas educacionais, sendo tipicamente referida como uma capacidade de ordem superior (MACLURE, 1999, p. 247). Por outras palavras, um texto com baixa densidade lexical configura-se como mais dêitico, mais situado no *aqui-agora* do discurso, sendo comumente um texto com características da modalidade falada; por oposição, um texto que apresenta taxas mais altas de densidade lexical estará mais próximo da modalidade escrita.

Há diferentes formas de medição da densidade lexical, mas duas destas já se tornaram tradicionais na literatura. A primeira proposta vem de Ure (1971) e Halliday (1987), que se fundamentam na relação entre palavras lexicais, ou palavras de conteúdo, e palavras gramaticais, ou palavras funcionais. Essa relação expressa-se na forma de uma proporção. A técnica de Ure (1971, p. 445) consiste na obtenção da razão de palavras lexicais pelo total de palavras, o que é secundado por Stubbs (1996, p. 72), que acrescenta que a densidade deve ser expressa pela percentagem de ocorrências de palavras lexicais. Tome-se, para a ilustração da aplicação da medida, os excertos em 1 e 2, com as palavras lexicais 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)[3]

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).

No exemplo 1, há um conjunto de 16 palavras lexicais para um total de 52 palavras (das quais 36 são gramaticais), de que resulta uma razão de densidade lexical, se considerada a fórmula palavras lexicais/palavras totais, de 0,31. Pode-se dizer que 30,8% dos itens totais realizados são lexicais. Inversamente, do total de palavras empregues nesse excerto, 69,2% são palavras gramaticais. Quanto ao exemplo 2, que tem um total de 56 palavras, das quais 25 são lexicais e 31 são gramaticais, tem-se uma razão de 0,45, sendo possível afirmar que 44,6% das palavras equivalem a palavras lexicais e que 55,4% são gramaticais.

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

A segunda proposta de medição da densidade lexical é também de Halliday (1987, 2005, 2009), para quem a densidade se mede pela frequência média de palavras lexicais por orações, excluindo-se as orações encaixadas. O linguista sustenta que a oração é a unidade mais óbvia, mais natural para se avaliar a densidade, visto ser a maior unidade gramatical da língua, mas não exclui a possibilidade de se considerarem outras unidades, apresentando recentemente uma definição mais abrangente (2009, p. 75): "the quantity of lexicalized information packed into a given unit in the grammar". No entanto, Halliday afirma (2005, p. 168-169) que a aplicação do cálculo de densidade no escopo da oração, como ele propõe, é mais significativa quando na comparação da fala com a escrita, já que aquela, ainda segundo o linguista, tende a ser construída com mais orações do que esta. Por isso, não se utiliza esta forma de medição aqui, já que se trata de uma comparação entre textos unicamente da modalidade escrita. Acrescente-se que as medidas até este ponto referidas são convergentes em termos de resultados, como garantem Berman e Ravid (2009, p. 98), podendo a sua aplicação em simultâneo ao mesmo conjunto de dados resultar redundante.

Wolfe-Quintero, Inagaki e Kim (1998) listam várias outras técnicas de medição do desenvolvimento do léxico, como, por exemplo, a razão de nomes ou de verbos pelo total de palavras, apesar de referirem estas técnicas para a avaliação da diversidade, entendo-as como mais adequadas à avaliação da densidade, já que envolvem a especialidade categorial, tornando-se, portanto, complementares às medidas clássicas de Ure (1971) e Halliday (1987). Deste modo, torna-se possível falar em densidade nominal (razão de nomes pelo total de palavras) ou em densidade verbal (razão de verbos pelo total de palavras). Neste trabalho, portanto, utiliza-se primeiramente a medição clássica – razão de palavras lexicais pelo total de palavras –, (que nomeio de densidade lexical global), e, secundariamente, aplicam-se as medidas por classes de palavras (ou densidade lexical por especialidade).

As discussões sobre como calcular a densidade não se têm centrado somente em questões de natureza matemática, mas também em questões de natureza linguística. Dizem respeito à diferenciação entre palavras lexicais e palavras gramaticais. Ure (1971, p. 445), por exemplo, inclui, como palavras lexicais, ou palavras com propriedades lexicais – em oposição às palavras com propriedades gramaticais – nomes, verbos e adjetivos, indistintamente.

Strömqvist et al. (2001, p. 48) admitem como palavras lexicais nomes, adjetivos e verbos, mas excluem os verbos auxiliares e os verbos copulativos. Há ainda autores que optam por descrever ocorrências de classes específicas, como fazem Ravid e Berman (2006), que avaliam as ocorrências de nomes num grupo de textos falados e escritos em registros narrativos e não narrativos.

Como observa Halliday (1989, p. 61), entre palavras lexicais e palavras gramaticais existe uma relação escalar, não existindo uma divisão nítida entre umas e outras, mas é possível instituí-la em favor da coerência metodológica. Neste trabalho, consideram-se as recomendações de Mendes (2013, p. 258), que, aceitando também que há uma relação escalar entre palavras lexicais e palavras gramaticais, admite como itens mais próximos do polo lexical os seguintes: nomes, adjetivos (incluindo numerais ordinais), verbos plenos e advérbios terminados em -mente.

2 Métodos e *Corpus*

A fim de avaliar a correlação entre a densidade 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, de quatro escolas públicas da Grande Lisboa, alunos regularmente matriculados no 5º ano, a representar o desenvolvimento lexical ocorrido no 1º ciclo; alunos matriculados no 7º ano, a representar o desenvolvimento lexical no 2º ciclo; e alunos matriculados no 10º ano, a representar o desenvolvimento lexical ocorrido no 3º ciclo. O acesso às escolas, a aplicação dos estímulos, a obtenção e o consequente manuseio para fins investigativos dos textos realizou-se sob uma perspectiva ético-legal, pelo que se seguiram os procedimentos para investigação em meio escolar regidos pelos termos do Despacho Nº 15847/2007, publicado no DR 2ª série nº 140, de 23 de julho, e concebido especificamente para regular a aplicação de inquéritos em meio escolar. Fundamentado por este despacho, o acesso às escolas foi aprovado pela Direção Geral de Inovação e Desenvolvimento Curricular (DGIDC), sob o número de processo 0260700001.

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 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, foram corrigidos desvios de natureza gráfica, tais como acentuação, capitalização ou indentação. As informações gerais sobre os participantes e sobre o *corpus*, intitulado CODES[4] (MARTINS, 2015), descrevem-se no Quadro 1, a seguir:

Quadro 1: Informações gerais sobre os participantes e sobre o *corpus* (CODES)

	5º	**7º**	**10º**	**Total**
Idade (M, DP)	10,19 (0,402)	12,33 (0,701)	15,16 (0,370)	—
Sexo				
feminino (%)	18 (69,2%)	25 (54,3%)	29 (58%)	72 (59%)
masculino (%)	8 (30,8%)	21 (45,7%)	21(42%)	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)	—
Nº total de palavras	8.586	15.239	19.499	4.324
Nº total de textos	52	92	100	244

Fonte: CODES.

[4] O *corpus* CODES está atualmente disponível para consulta em <http://alfclul.clul.ul.pt/CQPweb/codes/>.

Os textos foram anotados morfossintaticamente com o MBT (*memory-based tagger*) (DAELEMANS et al., 1996) e lematizados com o MBLEM (VAN DEN BOSCH; DAELEMANS, 1999), adaptados para o português (GÉNÉREUX; HENDRICKX; MENDES, 2012). Para a extração de listas de palavras lexicais e gramaticais 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[5].

No processador CQP, após definir como contexto apenas uma palavra (set Context 1) e solicitar a exibição dos atributos posicionais 'pos' e 'lemma' (show +pos; show +lemma) por texto (set PrintStructures "text_id"), foi extraída uma lista de palavras por ano e por registro textual. Os dados, delimitados por tabulação, foram, de seguida, importados para o programa Numbers, onde foi possível configurar as ocorrências, por texto, de todas as categorias POS, sendo possível, de seguida, listar as ocorrências de palavras pertencentes às classes lexical e gramatical, diferenciadamente. Como já referido, a classe lexical é constituída por nomes, verbos plenos, adjetivos e advérbios terminados em -mente (MENDES, 2013). No CQP, a classe 'nome' foi selecionada a partir das seguintes etiquetas: CN (nome comum), PNM (nome próprio), MTH (mês), WD (dia da semana) e STT (títulos sociais – como 'Presidente', 'dr.', 'prof.', etc.). Na composição da categoria 'verbo pleno', incluíram-se as etiquetas V (verbo), GER (gerúndio), INF (infinitivo), PPA (particípio não constituinte de tempo composto), e excluíram-se as categorias GERAUX (gerúndio como verbo auxiliar), INFAUX (infinitivo como verbo auxiliar) e PPT (particípio em tempos compostos)[6]. Também foram excluídos os verbos de ligação, ou verbos copulativos. Em virtude de o anotador MBT (DAELEMANS et al..valor médio obtido é de 44,47;

[5] As plataformas CLAN e IMS Open Corpus Workbench, bem como os manuais, estão gratuitamente disponíveis para *download*, respectivamente, em: <http://childes.psy.cmu.edu/> e <http://cwb.sourceforge.net/>.

[6] Para mais informações, consulte-se o *Manual for the CRPC CQPweb interface* (GÉNÉREUX; HENDRICKX; MENDES, 2011), disponível em: <http://alfclul.clul.ul.pt/CQPweb/doc/CQPweb_ma- nual.v1.pdf>.

como proposta por Jean , 1996) não ser treinado para identificar, em português, verbos de ligação, recorreu-se ao anotador *on-line* Palavras (BICK, 2000)[7].

Considera-se, para todos os efeitos estatísticos, o ano escolar (5º, 7º e 10º) 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 densidade, consideraram-se os valores das razões obtidas. 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, todos os dados 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 e" 0,05. Constatou-se que as razões obtidas, 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.

3 Resultados

No Quadro 2, apresenta-se a frequência relativa de ocorrências de palavras lexicais (*lex*) e de palavras gramaticais (*gram*), acompanhados dos respectivos desvios-padrão, enquanto itens constituintes necessários à obtenção da razão da densidade:

[7] Disponível em <http://beta.visl.sdu.dk/visl/pt/parsing/automatic/parse.php>.

Quadro 2: Frequência média (e desvio-padrão), em números percentuais, de palavras lexicais (lex) e de palavras gramaticais (gram) identificadas nos registros narrativo e argumentativo

Registro	Ano	N	lex (DP)		gram (DP)	
Narrativo	5º	26	44,47	4,10	55,53	4,10
	7º	46	43,93	4,93	56,07	4,93
	10º	50	43,94	3,80	56,06	3,80
Argumentativo	5º	26	45,53	4,94	54,47	4,94
	7º	46	47,23	4,13	52,77	4,13
	10º	50	47,35	4,47	52,65	4,47

Como demonstrado na tabela, o número médio de palavras lexicais no registro narrativo diminui do 5º para o 7º ano e eleva-se deste para o 10º ano, enquanto o número médio de palavras gramaticais cresce linearmente de ano a ano. No registro argumentativo, acontece um processo contrário, em que o número médio de palavras lexicais cresce de ano a ano, mas o número médio de palavras gramaticais decai do 5º para o 7º ano, elevando-se deste para o 10º ano.

Quanto à razão da densidade global identificada no registro narrativo, os valores percentuais médios decrescem do 5º para o 7º ano e mantêm-se aparentemente estáveis até ao 10º ano. No 5º ano, o valor médio obtido é de 44,47; no 7º ano, o valor médio é de 43,93; e, no 10º ano, o valor médio é de 43,94. Quanto ao registro argumentativo, o movimento do valor percentual da densidade é de ascensão do 5º ao 7º ano, e de ligeira diminuição deste ao 10º ano. Assim, no 5º ano, o valor médio da densidade é 45,16; no 7º ano, o valor médio é de 47,00; e, no 10º ano, o valor médio é de 46,85. No Gráfico 1, é possível ilustrar o movimento dos valores aqui apresentados:

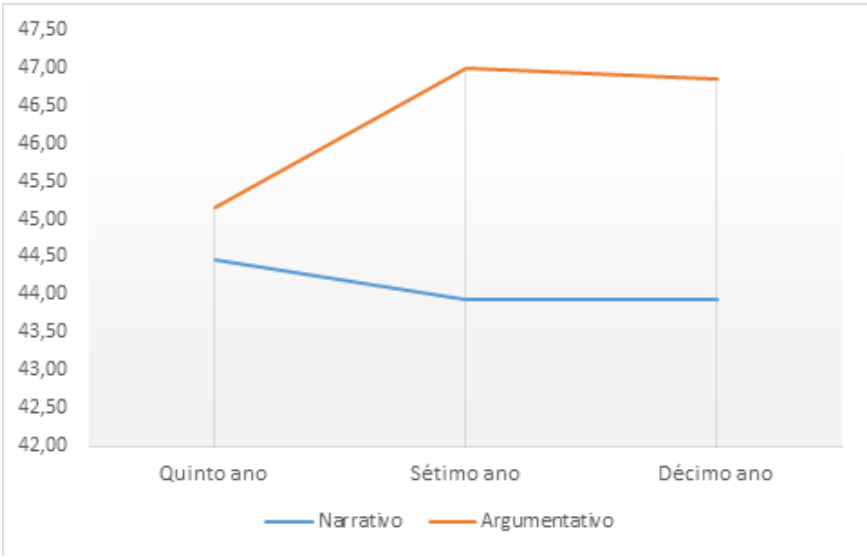

Gráfico 1: Distribuição dos valores médios da densidade lexical global, em valores percentuais, nos registros narrativo e argumentativo ao longo dos anos escolares

O teste estatístico, quanto aos textos narrativos, indica que não há correlação[8] (r=-0,031; p=0,736[9]) entre a densidade lexical global e a progressão escolar. Quer isto dizer que, independentemente do estádio em que se encontra o aluno na escola, a proporção de palavras lexicais em relação ao total de palavras mantém-se inalterada. Quanto aos textos de registro argumentativo, os testes de correlação indicam que também não existe correlação (r=0,142; p=0,120) entre a densidade lexical global e a progressão escolar, verificando-se situação semelhante ao que acontece nos textos de registro narrativo.

[8] 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.

[9] Correlação significativa no nível 0,01 (2 extremidades).

Os gráficos abaixo ilustram como se comportam individualmente as classes de palavras que constituem a categoria *lex* ao longo da progressão escolar, respectivamente nos textos narrativos e nos textos argumentativos. Os valores apresentados são percentuais em relação ao total de ocorrências de *tokens*:

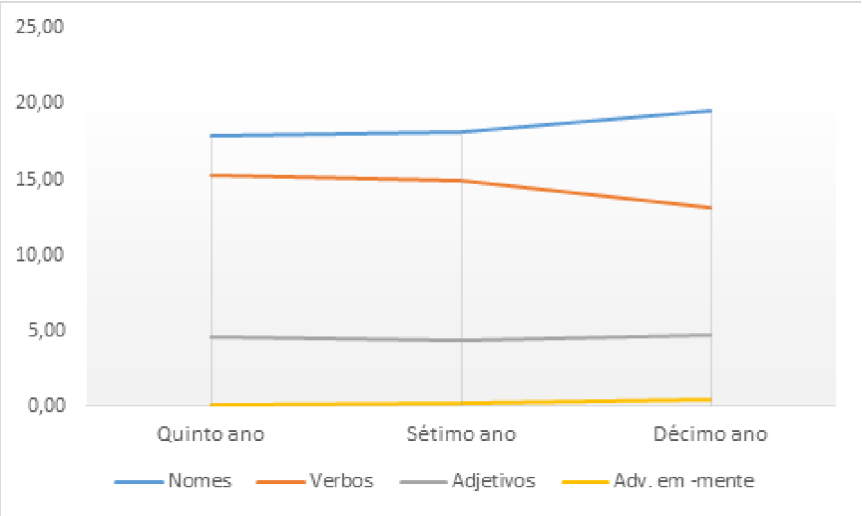

Gráfico 2: Distribuição dos valores médios de ocorrências, em percentual, de nomes comuns, verbos plenos, adjetivos e advérbios terminados em -mente no registro narrativo ao longo dos anos escolares

O Gráfico 2, que trata das ocorrências das classes de palavras lexicais nos textos narrativos, permite justificar a ausência de correlação entre a densidade lexical global e a progressão escolar, já que as percentagens de nomes, verbos, adjetivos e advérbios terminados em -mente mostram-se de uso homogêneo de ano para ano.

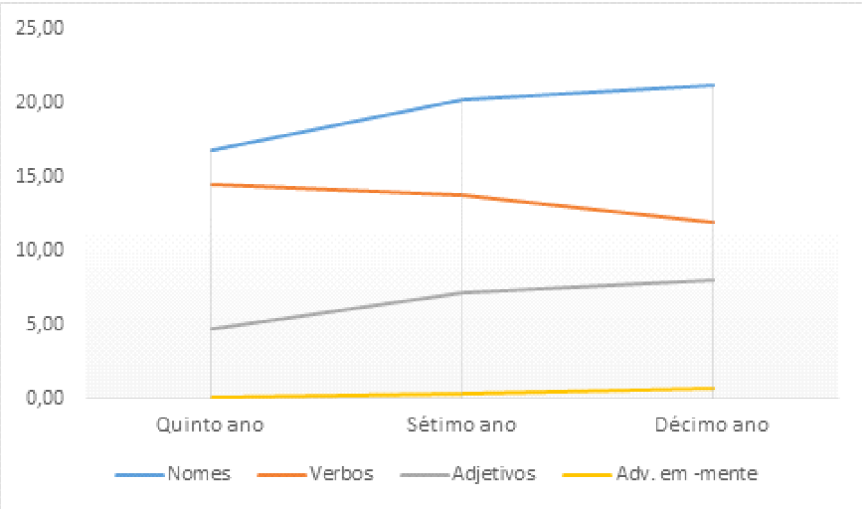

Gráfico 3: Distribuição dos valores médios de ocorrências, em percentual, de nomes comuns, verbos plenos, adjetivos e advérbios terminados em -mente no registro argumentativo ao longo dos anos escolares

A partir do Gráfico 3, que descreve esta medida nos textos argumentativos, pode-se inferir que a ausência de correlação entre a densidade lexical global e a progressão escolar justifica-se nos valores gradualmente opostos das ocorrências de nomes e verbos, que, a partir do 7º ano, se distanciam. No 5º ano, os verbos são mais recorrentes do que os nomes; a partir do 7º ano, passa-se a uma situação diferente, em que os nomes são mais acionados do que os verbos. Aparentemente, a percentagem de advérbios terminados em -mente mantém-se também inalterada ao longo dos anos, mas a de adjetivos mostra-se crescente de ano para ano. Contudo, a leitura dos gráficos acima não permite perceber, com precisão matemática, como se comportam as classes lexicais ao longo da progressão escolar. Deste modo, realizaram-se os testes estatísticos diferenciadamente para cada classe a fim de se obter dados sobre a densidade por especialização (nominal, verbal, adjetival e adverbial).

Sobre o emprego de nomes no registro narrativo, o teste de correlação aponta para uma correlação positiva fraca ($r=0,240$; $p=0,008$) entre a

densidade nominal e a progressão escolar. A análise da variância confirma a existência de diferenças estatisticamente significativas entre os anos de escolaridade (F(4,834); p=0,010[10]), com o teste *post-hoc* de Tukey a mostrar que essas diferenças se localizam apenas entre o 7º e o 10º ano (p=0,011), não se manifestando nem entre o 5º e o 7º ano (p=0,924), nem entre o 5º e o 10º (p=0,093). A densidade adjetival e a progressão escolar mantêm uma correlação positiva fraca (r=0,234; p=0,009), corroborada pela análise da variância (F(4,819); p=0,010). No entanto, o teste *post-hoc* indica que só há diferenças entre o 7º e o 10º ano (p=0,010), não existindo nem entre o 5º e o 7º (p=0,875), nem entre o 5º e o 10º (p=0,113). Ainda quanto aos textos narrativos, não se identificou correlação entre a ocorrência de verbos e a progressão escolar (r=0,129; p=0,156). Por fim, foi identificada uma correlação positiva moderada (r=0,404; p=0,000) entre a densidade adverbial (-mente) e a progressão escolar, corroborada pela análise da variância (F(12,072; p=0,000). O teste de Tukey mostra que não há diferença significativa entre o 5º e o 7º ano (p=0,580), havendo diferenças entre o 5º e o 10º ano (p=0,000) e entre o 7º e o 10º (p=0,000).

Quanto aos textos de registro argumentativo, identificou-se uma correlação positiva moderada (r=418; p=0,000) entre a densidade nominal e a progressão escolar, confirmada pelas diferenças encontradas na análise da variância (F(12,661); p=0,000). As diferenças entre os anos, pelo *post-hoc* de Tukey, veem-se entre o 5º e o 10º ano (p=0,000) e entre o 7º e o 10º ano (p=0,001), não se identificando diferenças entre o 5º e o 7º (p=0,295). A densidade adjetiva nos textos argumentativos e a progressão escolar mantêm entre si uma correlação positiva moderada (r=0,524; p=0,000). A análise de variância confirma a existência de diferenças (F(22,270); p=0,000), localizando-se estas diferenças na comparação entre todos os anos: entre o 5º e o 7º (p=0,009), entre o 7º e o 10º (p=0,000) e entre o 5º e o 10º (p=0,000). Ainda quanto aos textos argumentativos, pela aplicação dos testes estatísticos, não há correlação entre a ocorrência de verbos e a progressão escolar (r=0,012; p=0,893). Entre a densidade adverbial (-mente) e a progressão escolar identificou-se uma correlação positiva fraca (r=0,347;

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

p=0,000). A análise da variância confirma a diferença entre os grupos estudados (F(8,143); p=0,000), existindo especificamente diferenças entre o 5º e o 10º ano (p=0,001) e entre o 7º e o 10º ano (p=0,019), não acontecendo o mesmo entre o 5º e o 7º (p=0,303).

4 Discussão

Quanto à correlação entre a densidade lexical global (razão entre a ocorrência de palavras lexicais e o total de palavras, multiplicada por cem), e a progressão do 5º ao 10º ano da escolaridade, tanto nos textos narrativos como nos argumentativos, como se viu, não há movimentos de mudança significativos. Este resultado contraria a expectativa de que, na progressão escolar, há um adensamento lexical da escrita e, consequentemente, refuta a primeira hipótese de trabalho, segundo a qual os textos dos alunos mais velhos são mais densos lexicalmente do que os textos dos alunos mais novos, pensamento que se vê, por exemplo, em Schleppegrell (2004, p. 108), que afirma que quanto mais se avança nos anos escolares mais os alunos se tornam capazes de condensar informação através de palavras lexicais. Este resultado também anula, em princípio, uma possível percepção de que a escrita escolar pode ser tratada como uma faceta da escrita acadêmica, como considera Jisa (2004, p. 135). Contudo, Biber (2006, p. 14) destaca que a escrita acadêmica se caracteriza pela prevalência de nomes e pela ocorrência menos saliente de verbos, resultados que encontram eco nos textos aqui estudados quando se considera a densidade por classe, sendo identificadas correlações positivas, em ambos os registros, entre o emprego de nomes, adjetivos e advérbios em -mente e a progressão escolar, e nenhuma correlação entre o emprego de verbos e a progressão escolar.

Berman e Ravid (2009), num estudo experimental que considera participantes em idade escolar (dos nove aos dezenove anos) monolíngues de hebreu ou de inglês, relatam, por exemplo, que a taxa de densidade lexical, em particular de adjetivos, cresce significativamente em função da idade e em função do registro. Mais especificamente, os textos produzidos pelos participantes mais velhos, em idade escolar mais avançada, apresentam maiores taxas de densidade do que os textos dos mais novos. Ravid (2004) avalia a densidade global e por especialidade num estudo que compara a

complexidade linguística em textos expositivos e narrativos escritos por alunos hebreus monolíngues situados no 4º, 7º e 11º anos de escolaridade e constata que a densidade nominal cresce significativamente de ano a ano.

Os resultados dos testes estatísticos aplicados ao *corpus* aqui investigado, em que não se identificou correlação nenhuma, refutam consequentemente a segunda hipótese de trabalho, a de que a densidade lexical, considerada holisticamente, é maior nos textos narrativos, já que, como alguns estudos sugerem (BEREITER; SCARDAMALIA, 1982; CHISTIE, 1986), os alunos, ao longo da sua vida escolar, experimentam muitas dificuldades na escrita de textos não-narrativos, tornando-se estes tipos de textos grandes desafios, inclusivamente na seleção lexical, pois, mesmo sendo dependentes de um dado tema, habitualmente requerem um léxico mais especializado do que os textos narrativos. A refutação da hipótese secundária, se considerada a densidade por classe, mantém-se, desta feita não pela ausência de correlação, mas sim, como se verificou, pela força da correlação da densidade nominal, verbal e adjetival no registro narrativo, que é sempre menos expressiva do que a força no registro argumentativo. Semelhantemente, os resultados de Ravid (2004) apontam para um crescimento mais acentuado da densidade nominal e adjetival nos textos argumentativos. Berman e Ravid (2009), sobre a densidade adjetival, concluem que o crescimento é mais significativo nos textos expositivos do que o que se encontra nos textos narrativos. Para o português europeu, não foram identificados estudos sobre o desenvolvimento linguístico que se propusessem aferir a densidade lexical em textos escolares pela mesma métrica que aqui se utiliza.

Os resultados de Ravid (2004), 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 produzidos por crianças e jovens em idade escolar, por serem adquiridos ainda nos anos iniciais da escolaridade e por serem aparentemente mais bem escritos, se mostram 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 e das suas especificidades lexicais.

No registro argumentativo, ainda quanto aos resultados da medição da densidade por classe, observou-se que, à exceção do adensamento do léxico nominal e adjetival, que se torna significativo já entre o 5º e o 7º ano (e consequentemente entre o 5º e o 10º), em todos os outros casos de densidade por classe, as diferenças entre os grupos são significativas apenas entre o 7º e o 10º (ou entre o 5º e o 10º). Nos textos de registro narrativo, todas as diferenças identificadas igualmente ocorrem apenas entre entre o 7º e o 10º ano (ou entre o 5º e o 10º).

Considerações Finais

O estudo aqui reportado teve por objetivo central examinar a correlação entre um indicador de desenvolvimento lexical, nomeadamente a densidade, 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.

Embora os resultados dos testes de correlação aplicados aos indicadores de desenvolvimento lexical tenham revelado cenários de desenvolvimento estatisticamente semelhantes nos registros narrativos e argumentativos comparativamente, o exame da força da correlação demonstrou que a densidade nominal é bem mais expressiva nos textos argumentativos do que nos textos narrativos, permitindo concluir que há da parte dos alunos investigados alguma consciência de registro (e, de novo, de padrão linguístico), ou alguma consciência de que, para narrar e argumentar, enquanto variedades linguísticas escolares, são mobilizadas configurações de uso da língua distintas para cada variedade.

Esta conclusão não é de todo surpreendente, dado que, como lembra Reppen (2007, p. 156), a consciência de registro se torna evidente já a partir do 3º ano de escolaridade, mas, ainda de acordo com Reppen, é preciso considerar que esta tomada de consciência não se trata de um fenômeno estanque, já que continua a desenvolver-se nos anos de escolaridade seguintes, razão por que deve ser contínua e sistematicamente estimulada por meio de

franca exposição a uma diversidade alargada de configurações linguísticas próprias de determinados registros. Neste sentido, as orientações legais para o ensino de português, em particular as que se veem no Programa de Português (PORTUGAL, 2009), contemplam bem esta questão ao referir a necessidade de se constituir para cada ciclo escolar um "*corpus* textual", definido como um "conjunto alargado de objectos textuais que hão-de estar presentes na aula de Português, em diversos suportes, destinados ao desenvolvimento das competências específicas quer no modo oral, quer no modo escrito" (PORTUGAL, 2009, p. 100)[11].

Referências

BEREITER, C.; SCARDAMALIA, M. From conversation to composition: The role of instruction in a developmental process. In: GLASER, R. (Ed.). *Advances in instructional psychology*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1982. p. 73-96.

BERMAN, R. A. 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.; 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. *University language:* A corpus-based study of spoken and written registers. Amsterdã; Filadélfia: John Benjamins, 2006.

BICK, E. *The parsing system Palavras* – automatic grammatical analysis of Portuguese in a constraint grammar framework. 2000. Aarhus: Aarhus University, 2000.

[11] No Programa de Português mais recentemente homologado (PORTUGAL, 2015), já não se veem estas orientações explicitamente.

CHRISTIE, F. Writing in schools: Generic structures as ways of meaning. In: COUTURE, B. (Ed.). *Functional approaches to writing:* Research perspectives. Londres: Frances Pinter, 1986. p. 221-240.

COLOMBI, M. C. Academic language development in latino student's writing. In: SCHLEPPEGRELL, M. J.; COLOMBI, M. C. (Ed.). *Developing advanced literacy in first and second languages*. Mahwah: Lawrence Erlbaum Associates, 2002. p. 67-86.

DAELEMANS, W. et al. MBT: A memory-based part of speech tagger generator. In: WORKSHOP ON VERY LARGE CORPORA, 4., 1996, Copenhagen. *Anais*… Copenhagen, 1996.

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.

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.

GÉNÉREUX, M.; HENDRICKX, I.; MENDES, A. A large Portuguese corpus cleaning and preprocessing. In: COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE. Proceedings of the 10th International Conference PROPOR1012, 2012, Berlin, Heidelberg. *Anais*… Berlin; Heidelberg: Springer-Verlag, 2012.

HALLIDAY, M. A. K. Spoken and written modes of meaning. In: HOROWITZ, R.; SAMUELS, S. J. (Ed.). *Comprehending oral and written language*. Orlando: Academic Press, 1987. p. 55-82.

HALLIDAY, M. A. K. The spoken language corpus: A foundation for grammatical theory. In: WEBSTER, J. J. (Ed.). *Computational and quantitative studies*. Londres; Nova Iorque: Continuum, 2005. p. 157-190.

HALLIDAY, M. A. K. Methods – techniques – problems. In: HALLIDAY, M. A. K.; WEBSTER, J. J. (Ed.). *Continuum companion to systemic functional linguistics*. Londres; Nova Iorque: Continuum, 2009. p. 59-86.

HALLIDAY, M. A. K.; HASAN, R. *Language, text and context:* Aspects of language in a social-semiotic perspective. Oxford: Oxford University Press, 1989.

JISA, H. Growing into academic French. In: BERMAN, R. A. (Ed.). *Language development across childhood and adolescence*. Amsterdã; Filadélfia: John Benjamins, 2004. p. 135-162.

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

MACLURE, M. Home and school language. In: SPOLSKI, B. (Ed.). *Concise encyclopedia of educational linguistics*. Oxford: Pergamon; Elsevier, 1999. p. 202-203.

MARTINS, M. *Corpus desenvolvimental:* Codes. Lisboa: Centro de Linguística da Universidade de Lisboa (CLUL), 2015. Disponível em: <http://alfclul.clul.ul.pt/CQPweb/codes2016/>. Acesso em: 03 abr. 2016.

MARTINS, M. A diversidade lexical na escrita de textos escolares. *Fórum Linguístico*, v. 13, n. 1, p. 1068-1082, 2016.

MENDES, A. Organização textual e articulação de orações. In: RAPOSO, E. B. P. et al. (Ed.). *Gramática do português*. Lisboa: Fundação Calouste Gulbenkian, 2013. p. 1691-1758.

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

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

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

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. p. 337-355.

RAVID, D.; BERMAN, R. A. Information density in the development of spoken and written narratives in English and Hebrew. *Discourse Processes*, v. 41, n. 2, p. 117-149, 2006.

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

REPPEN, R. First language and second language writing development of elementary students. In: KAWAGUCHI, Y. et al. (Ed.). *Corpus-based perspectives in linguistics*. Amsterdã; Filadélfia: John Benjamins, 2007. p. 147-168.

SCHLEPPEGRELL, M. J. *The language of schooling:* A functional linguistics perspective. Mahwah; Londres: Lawrence Erlbaum, 2004.

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.

STUBBS, M. *Text and corpus analysis:* Computer assisted studies of language and culture. Oxford: Blackwell, 1996.

URE, J. Lexical density and register differentiation. In: PERREN, G. E.; TRIM, I. L. M. (Ed.). *Applications of linguistics*. Cambridge: Cambridge University Press, 1971. p. 443-452.

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.

WOLFE-QUINTERO, K.; INAGAKI, S.; KIM, H. Y. *Second language development in writing:* Measures of fluency, accuracy, and complexity. Honolulu: University of Hawai'i at Mãnoa, 1998.

WRAY, D.; MEDWELL, J. *Progression in writing and the Northern Ireland levels for writing*. Warwick: University of Warwick, 2006.

Recebido em: 04/04/2016
Aceito em: 03/09/2016