revista HORIZONTES DE LINGUÍSTICA APLICADA

e-ISSN 2237-0951

ARTIGO ORIGINAL

Uma análise de frequência lexical em materiais didáticos de Inglês do e-Tec Idiomas

A lexical frequency analysis in English teaching materials of the e-Tec Idiomas Program

Camila De Bona[1], Luiz Carlos Schwindt[2], Daniel Vinícius Böch[3]

1 Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense —camilabona@ifsul.edu.br
2 Universidade Federal do Rio Grande do Sul — schwindt@ufrgs.br
3 Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense —danielboch.ss024@academico.ifsul.edu.br

Como citar o artigo
DE BONA, C.; SCHWINDT, L. C.; BÖCH, D. V. Uma análise de frequência lexical em materiais didáticos de inglês do e-tec idiomas. *Horizontes de Linguística Aplicada*, ano *25, n.* 1, AG4, 2026.

RESUMO

Pesquisas recentes indicam que palavras de alta frequência devem ser priorizadas no aprendizado de vocabulário em língua estrangeira. Contudo, as frequências lexicais em livros didáticos, principal recurso dos aprendizes, nem sempre correspondem às observadas em grandes *corpora* (Nation, 2011; Sakata, 2019). Este estudo analisa a frequência de verbos, substantivos e adjetivos nas apostilas de Inglês Módulo I do e-Tec Idiomas, em comparação ao *Corpus of Contemporary American English* (COCA *Corpus*). O projeto é relevante para o ensino de línguas, pois, ao examinar um material gratuito usado em instituições públicas, pode propor melhorias que favoreçam a aprendizagem, alinhando a frequência lexical no material ao uso real da língua. O referencial teórico baseia-se em Nation (2006; 2011), Criado e Sánchez (2012) e Norberg e Nordlund (2018). A metodologia envolveu a coleta manual de palavras e a análise da correspondência entre o material e o *corpus* de referência. Os resultados mostram que 85,59% dos verbos pertencem aos 1.000 lemas mais frequentes do inglês, enquanto substantivos e adjetivos alcançam 45,66% e 45,44%. O índice *type/token* de 0,246 revela-se adequado a materiais de nível básico. Por fim, palavras menos frequentes cumprem funções temáticas e gramaticais, atendendo aos objetivos do nível A1 do *Common European Framework of Reference for Languages* (CEFR) (Council of Europe, 2001).

Palavras-chave: Frequência lexical. Materiais didáticos. Ensino de Inglês. e-Tec Idiomas.

ABSTRACT

Recent research indicates that high-frequency words should be prioritized in foreign language vocabulary learning. However, lexical frequencies in textbooks, the main resource for learners, do not always match those observed in large *corpora* (Nation, 2011; Sakata, 2019). This study analyzes the frequency of verbs, nouns, and adjectives in the English Module I booklets of the e-Tec Idiomas program, compared to the *Corpus* of Contemporary American English (COCA *Corpus*). The project is relevant to language teaching because, by examining a free educational material used in public institutions, it can propose improvements that support learning by aligning lexical frequency in the material with actual language use. The theoretical framework is based on Nation (2006; 2011), Criado and Sánchez (2012),

Fonte de financiamento: Propesp IFSul – Edital 09/2022
Conflito de interesse: Os autores declaram não haver
Recebido em: 31 Ago 2025. Aceito em: 05 Fev 2026.

and Norberg and Nordlund (2018). The methodology involved manual word collection and analysis of the correspondence between the material and the reference *corpus*. Results show that 85.59% of verbs belong to the 1,000 most frequent English lemmas, while nouns and adjectives reach 45.66% and 45.44%, respectively. The type/token ratio of 0.246 is suitable for basic-level material. Finally, less frequent words fulfill thematic and grammatical functions, meeting Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) A1 level objectives.

Keywords: Lexical frequency. Teaching materials. English language. e-Tec Idiomas.

1 INTRODUÇÃO

Considerando que o vocabulário de uma pessoa é composto de palavras de alta e baixa frequência, diversas pesquisas no âmbito da Linguística destacam a importância das palavras de alta frequência no processo de aprendizagem de uma língua adicional. Nation (2006; 2011) aponta que os professores devem lidar com palavras de alta frequência de maneira muito diferente da forma como lidam com palavras de baixa frequência. A ideia por trás dessa distinção é que as palavras de alta frequência formam um grupo de palavras relativamente pequeno e muito útil, que são importantes independentemente do uso que seja feito delas. Como cada palavra nesse grupo é frequente, os estudantes em geral têm um retorno bastante significativo ao aprendê-las. As palavras de baixa frequência, por outro lado, consistem de dezenas de milhares de palavras que ocorrem muito raramente, são comumente restritas a certas áreas temáticas e, portanto, não merecem atenção substancial na sala de aula.

Nation (2011) defende que estudantes de línguas adicionais tenham como trabalho principal aprender as palavras de alta frequência do idioma, o que se torna mais fácil se eles: (i) encontrarem essas palavras com frequência; (ii) concentrarem-se deliberadamente nelas; e (iii) não se distraírem por um número substancial de palavras de baixa frequência. Tendo em mente que livros didáticos são amplamente utilizados no processo de ensino e aprendizagem de inglês, muitos professores acreditam que esses materiais têm um planejamento razoável para ensino de vocabulário. No entanto, diversas análises de frequência lexical em livros didáticos (Criado; Sánchez, 2012; Norberg; Nordlund, 2018; Sakata, 2019) apontam que, infelizmente, esse nem sempre é o caso.

Com isso em vista, nosso objetivo neste trabalho é analisar a frequência lexical de verbos, nomes e adjetivos presentes nas apostilas de Inglês Módulo I de um programa de ensino de línguas a distância, o e-Tec Idiomas, comparativamente a um *corpus* representativo dessa língua, o *Corpus of Contemporary American English* (COCA *Corpus*), o qual nos permite identificar a frequência das palavras no uso real da língua. Mais especificamente, temos o intuito de: (i) descrever a distribuição dos itens lexicais por faixas de frequência, verificando quantas palavras de cada classe gramatical presentes no material didático estão entre as 1.000 mais frequentes da língua; e (ii) verificar o número de verbos, substantivos e adjetivos diferentes apresentados aos estudantes nas apostilas e a quantidade de vezes que essas palavras são repetidas.

O e-Tec Idiomas é concebido e produzido pela Rede Federal de Educação, Ciência e Tecnologia (RFEPCT), com o intuito de suprir uma carência de oferta de línguas adicionais da própria Rede. Os materiais foram desenvolvidos, inicialmente, com o objetivo de "*desenvolver os programas de mobilidade e capacitação dos estudantes para o mundo do trabalho*" (e-Tec Idiomas Sem Fronteiras, [s.d.]). Atualmente, o Instituto Federal Sul-rio-grandense (IFSul) oferece, semestralmente, cursos de Formação Inicial e Continuada (FIC) Educação a Distância (EaD), massivos e sem tutoria, fazendo uso dos materiais do e-Tec Idiomas. No contexto dessa demanda particular é que se enquadra este estudo, pois, considerando o elevado número de pessoas que utilizam esse material didático sem auxílio direto de um professor, uma análise linguística em termos de frequência lexical torna-se pertinente. Dessa forma, nossa hipótese para esta pesquisa centra-se na ideia de que, se tivermos uma correspondência positiva entre

a frequência lexical e a frequência de ocorrência no material, a aprendizagem será favorecida – ainda que não seja objetivo deste estudo sua mensuração.

2 REFERENCIAL TEÓRICO

Pesquisas recentes sugerem que palavras de alta frequência devem ser priorizadas no processo de aprendizagem de vocabulário em uma língua estrangeira. No entanto, as frequências lexicais em livros didáticos, que são, em geral, o material mais utilizado por aprendizes, diferem daquelas encontradas em grandes *corpora* (Sakata, 2019).

Em artigo publicado em 2006, intitulado "How large a vocabulary is needed for reading and listening?", Paul Nation divide as palavras por meio de faixas de frequência, desenvolvendo 14 listas, cada uma com 1.000 palavras cada, com base nas informações encontradas no *corpus* de referência intitulado *British National* Corpus (BNC). O autor explica que a elaboração dessas listas teve como objetivo representar a faixa de maior frequência do vocabulário de um aprendiz, partindo do pressuposto de que tanto falantes nativos quanto não nativos adquirem vocabulário em grande parte de acordo com sua frequência e amplitude de uso, sendo as palavras mais frequentes e de maior alcance geralmente aprendidas antes daquelas menos frequentes e de alcance mais restrito (Nation, 2006).

Os dados examinados pelo pesquisador sugerem que a primeira faixa de frequência, ou seja, a primeira lista que contém as palavras mais frequentes, juntamente com nomes próprios, cobre entre 78% e 81% de texto escrito e cerca de 84% de texto falado. Já a quarta e a quinta faixas fornecem cerca de 3% de cobertura do texto escrito e 1,5% a 3% de cobertura do texto falado. Na Tabela 1, a seguir, vemos mais detalhadamente as porcentagens de cobertura de todas as faixas de frequência analisadas por Nation (2006).

Tabela 1. Alcance de cobertura por faixas de frequência.

Levels	Number of levels	Approximate written coverage (%)	Approximate spoken coverage (%)
1st 1,000	1	78—81	81—84
2nd 1,000	1	8—9	5—6
3rd 1,000	1	3—5	2—3
4th—5th 1,000	2	3	1.5—3
6th—9th 1,000	4	2	0.75—1
10th—14th 1,000	5	< 1	0.5
Proper nouns	1	2—4	1—1.5
Not in the lists	1	1—3	1

Fonte: Nation (2006, p. 79)

Percebe-se que, conforme a ordem das faixas avança e a frequência das palavras diminui, a cobertura, tanto de textos escritos quanto falados, decresce progressivamente.

Criado e Sánchez (2012) assumem que a repetição lexical é um fator de extrema relevância na aprendizagem de línguas. Os autores analisaram o léxico em dois livros didáticos de inglês, a fim de verificar se oportunidades de repetição lexical eram oferecidas. Os resultados dos pesquisadores mostraram que os dois livros são semelhantes no que diz respeito à quantidade de palavras utilizadas, mas diferem na ênfase da prática repetitiva.

Ao analisar o conteúdo lexical de sete livros didáticos de inglês utilizados nas escolas primárias da Suécia, Norberg e Nordlund (2018) mostraram que há variação considerável entre os livros, assim como que diz respeito à seleção e à quantidade de vocabulário como

no que tange à relação *type/token*[1]. Além disso, as pesquisadoras apontam que os livros analisados contêm uma grande proporção de palavras de baixa frequência, ou seja, de palavras pouco usadas na linguagem cotidiana. Ademais, de forma geral, a relação *type/token* nos livros estudados é alta, isto é, os estudantes apresentam poucas oportunidades de repetição e, por consequência, de consolidação do vocabulário estudado.

As pesquisadoras também analisaram separadamente as classes de palavras dos verbos, dos nomes e dos adjetivos. Seus resultados apontam que há correspondência entre a frequência de verbos nos livros didáticos e nos *corpora* de referência utilizados; no entanto, grandes diferenças no perfil lexical de adjetivos e substantivos indicam que a distribuição de palavras na produção da linguagem cotidiana não foi considerada na seleção de palavras para os livros.

O estudo de Norberg e Nordlund (2018) fornece evidências empíricas sobre a falta de critérios comuns quanto ao tamanho, à diversidade e ao perfil lexical em livros didáticos de Inglês usados nas escolas primárias suecas e, em dada medida, embasa este estudo. As autoras defendem que, uma vez que o livro é central em muitas atividades de ensino de línguas, análises de materiais didáticos com base na linguística de *corpus* devem ser empreendidas.

3 METODOLOGIA

O *corpus* deste estudo é o material didático de Inglês Módulo I do e-Tec Idiomas[2], que corresponde ao nível A1 (iniciante) do Quadro Europeu Comum de Referências para Línguas (QECR). Este primeiro módulo é composto por três apostilas de conteúdo, cada uma com 6 lições. Com isso em vista, foi realizada a extração manual das palavras de 18 lições, as quais apresentam seções de explicação linguística, diálogos, textos e seus respectivos glossários. Apresentamos, a seguir, Figura 1, a capa de uma das apostilas e uma página interna, que contém texto e glossário.

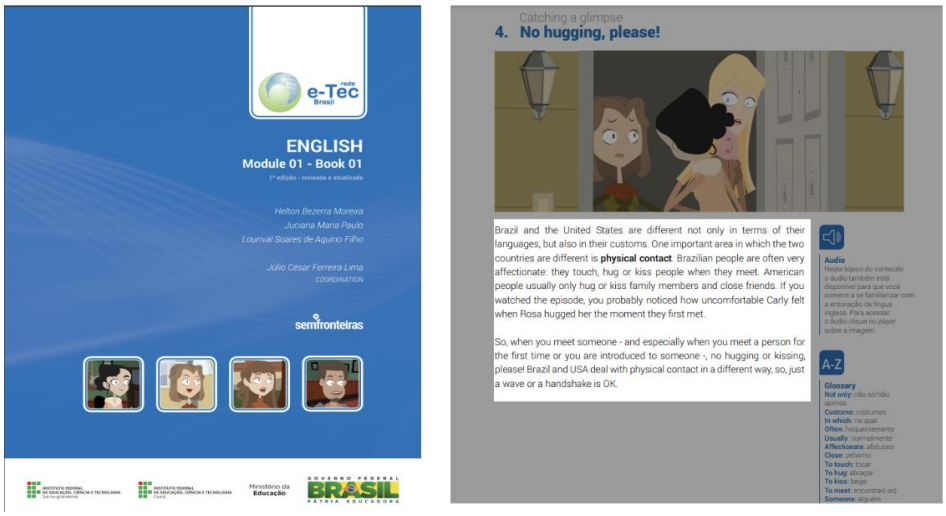

Figura 1. Capa e página interna da apostila de Módulo 1 do e-Tec Idiomas.
Fonte: Moreira *et al.* (2015).

[1] A relação *type/token*, também conhecida como *type-token ratio* (TTR), é utilizada como medida de diversidade lexical: enquanto *types* são as palavras únicas ou distintas em um texto, *tokens* referem-se às ocorrências dessas palavras, incluindo repetições. Ou seja, textos mais diversos têm uma TTR maior, enquanto textos mais repetitivos a têm menor.
[2] Os materiais podem ser acessados na íntegra em cpte.ifsul.edu.br.

Para verificar em que medida o vocabulário presente no material didático reflete o uso real da língua, cada palavra das apostilas foi comparada com sua respectiva ocorrência no COCA *Corpus*, um dos maiores *corpora* do inglês americano, contendo mais de um bilhão de palavras.

Faz-se importante salientar que este *corpus* permite estudos de lemas, *types* e *tokens*. Para fins de análise, consideramos lema a forma básica ou canônica de uma palavra, usada para representar, num amálgama, todas as suas flexões e variações gramaticais (por exemplo, *be* para *am, is, are, was, were, being, been, be*; *time* para *time, times*; *new* para *new, newer, newest*). Já um *type* refere-se às palavras diferentes, eliminando repetições, mas sem agrupar formas flexionadas ou variantes gramaticais (por exemplo, *be, is, are* são *types* diferentes, assim como *time* e *times*). Por fim, *token* é qualquer ocorrência individual de uma palavra, incluindo repetições e variações gramaticais: *tokens* representam cada instância em que uma palavra aparece.

Com isso em vista, nossos objetivos descritivos específicos neste trabalho são os que seguem:

Descrever a distribuição dos lemas de verbos, substantivos e adjetivos por faixas de frequência.

Verificar a relação *type/token*, ou seja, o número de verbos, substantivos e adjetivos diferentes apresentados aos estudantes em relação ao número total de itens de cada categoria e ao número total de palavras.

No que diz respeito ao primeiro objetivo, a versão gratuita do COCA *Corpus* disponibiliza informações sobre os 5.050 lemas mais frequentes da língua. Nesse sentido, considerando os estudos de Nation (2006) sobre a relevância dos 1.000 lemas mais frequentes, assim como as possibilidades de análise em nosso *corpus* de referência, nossa classificação por faixas de frequência foi a seguinte: faixa 1, os primeiros 1.000 lemas; faixa 2, entre 1.001 e 5.050; faixa 3, acima de 5.050.

Em relação ao segundo objetivo, como já discutido na seção de Referencial Teórico, a relação *type/token*, ou *type-token ratio* (TTR), é utilizada como medida de diversidade lexical. Divide-se o número de *types* (palavras únicas) pelo número de *tokens* (total de palavras), e seu resultado varia entre 0 e 1. Espera-se que livros didáticos de níveis mais avançados tenham uma TTR maior, por apresentarem mais diversidade lexical, enquanto materiais didáticos de nível básico a tenham menor, por serem mais repetitivos e apresentarem aos estudantes oportunidades de revisão e consolidação de vocabulário – idealmente, mais oportunidades de encontro com palavras mais frequentes.

Para realizar a análise quantitativa, aqui restrita à perspectiva descritiva, utilizamos a Plataforma R (R Core Team, 2024) a partir da interface R Studio.

4 RESULTADOS QUANTITATIVOS

4.1 Distribuição dos lemas por faixas de frequência

Em um primeiro momento, classificamos e quantificamos todas as palavras presentes na apostila, tanto lexicais quanto funcionais. Ao todo, tivemos um total de 21.043 palavras; deste número, 19.274 constam no COCA *Corpus* entre as 5.050 mais frequentes (Gráfico 1).

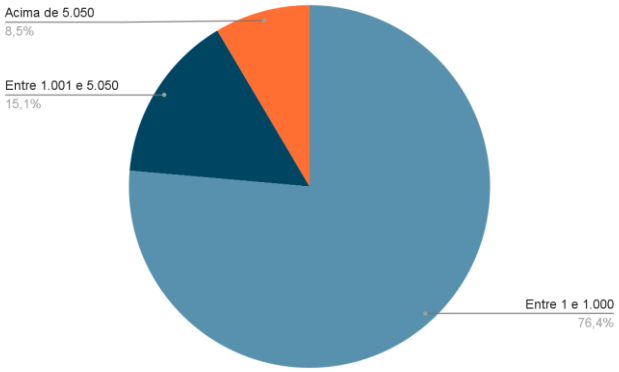

Gráfico 1. Distribuição de palavras funcionais e lexicais por faixas de frequência.
Fonte: elaboração dos autores.

Das 1.769 palavras acima de 5.050, a maioria são nomes próprios (de lugares e pessoas), interjeições (*jezz, hum, yeah*), advérbios (*comfortably, professionally*), numerais, principalmente os ordinais (*thirteenth, thousandth*), pronomes de tratamento (*Mr., Mrs., Ms.*), entre outros. Enquanto a classificação gramatical das palavras de 1 a 5.050 foi retirada do COCA *Corpus*, a classificação das palavras acima de 5.050 foi realizada manualmente. Após selecionar apenas verbos, substantivos e adjetivos, ficamos com 1.103 *tokens* que pertencem à faixa de frequência acima de 5.050.

A partir deste momento, nosso foco de análise centra-se nas palavras classificadas como verbos, substantivos e adjetivos, palavras portadoras de conteúdo essenciais para a comunicação. Considerando essas três classes gramaticais, temos um total de 10.604 ocorrências (*tokens*) e 1.977 lemas. A seguir, Tabela 2, apresentamos os dados divididos por classe.

Tabela 2. Número de lemas e *tokens*.

Verbos		Substantivos		Adjetivos	
Tokens	Lemas	Tokens	Lemas	Tokens	Lemas
4.462	384	4.663	1.185	1.479	408

Fonte: elaboração dos autores.

Dos 10.604 *tokens* extraídos do material didático, 6.620 estão entre os 1.000 mais frequentes, 2.881 na faixa entre 1.001 e 5.050, e 1.103 *tokens* pertencem à faixa de frequência acima de 5.050. Agora, vejamos essa divisão por classe de palavra (Tabela 3).

Tabela 3. Frequência de lemas.

Nível de Frequência do lema	Verbos (4.462 *tokens*)	Substantivos (4.663 *tokens*)	Adjetivos (1.479 *tokens*)
1 - 1.000	85,59% (3.819)	45,66% (2.129)	45,44% (672)
1.001 - 5.050	11,77% (525)	39,01% (1.819)	36,31% (537)
Acima de 5.050	2,64% (118)	15,33% (715)	18,26% (270)
1 - 1.000	85,59% (3.819)	45,66% (2.129)	45,44% (672)

Fonte: elaboração dos autores.

Como consta na Tabela 3, 85,59% dos verbos presentes no material didático do e-Tec Idiomas pertencem aos 1.000 lemas mais frequentes da língua inglesa, de acordo com o COCA

Corpus. Em relação aos substantivos e adjetivos, esse número é consideravelmente menor, perfazendo apenas 45,66% e 45,44%, respectivamente. No que diz respeito à faixa intermediária de frequência, enquanto temos 11,77% de verbos, essa porcentagem chega a 39,01% na classe dos substantivos e 36,31 na dos adjetivos. Em relação aos lemas que estão na faixa de frequência acima de 5.050, temos apenas 2,64% de verbos; já nas classes nominais, essa porcentagem sobe para 15,33% e 18,26%. Com isso em vista, considerando a importância dos lemas na primeira faixa de frequência, vemos que o resultado dos verbos é mais satisfatório que o resultado das classes nominais.

Uma maior porcentagem de verbos na primeira faixa de frequência comparativamente a substantivos e adjetivos também foi o resultado de Norberg e Nordlund (2018). No entanto, há uma diferença na análise: enquanto as pesquisadoras consideraram apenas verbos lexicais em seu estudo, nossa análise não apresenta divisão entre verbos auxiliares e lexicais (por exemplo, *do, did, have*), tendo em vista que o COCA *Corpus* não apresenta essa divisão na classificação dos seus lemas.

Na mesma linha de Norberg e Nordlund (2018), listamos os 10 lemas mais frequentes de verbos, substantivos e adjetivos do COCA *Corpus* e das apostilas de Módulo 1 do e-Tec Idiomas, com o intuito de comparar o perfil de vocabulário do material didático com o da língua inglesa de forma geral. Os lemas e seus números de ocorrência estão apresentados na Tabela 4.

Tabela 4. Os 10 lemas mais frequentes de verbos, substantivos e adjetivos no COCA Corpus e no e-Tec Idiomas.

Verbos				**Substantivos**				**Adjetivos**			
COCA *Corpus*		e-Tec Idiomas		COCA *Corpus*		e-Tec Idiomas		COCA *Corpus*		e-Tec Idiomas	
Be	32.394.756	**Be**	1.343	***Time***	2.018.725	***Time***	68	***Other***	1.539.952	***New***	44
Have	10.514.314	**Do**	314	***People***	1.800.205	***People***	55	***Good***	1.111.721	***Good***	38
Do	8.186.412	**Have**	241	***Year***	1.729.962	***Day***	55	***New***	1.017.175	***Great***	37
Say	4.096.416	**Go**	170	*Way*	1.260.011	***Year***	46	***Great***	696.589	*Nice*	35
Go	3.546.732	**Can**	140	*Thing*	1.202.004	*Park*	44	*Big*	600.364	***Old***	34
Get	3.347.615	Play	81	*Man*	1.091.176	*Student*	43	*High*	567.720	***Other***	28
Can	3.091.046	**Get**	74	***Day***	1.068.902	*College*	40	***Old***	537.424	*Cold*	23
Know	2.761.628	Take	68	*Life*	852.257	*Class*	39	*Bad*	426.558	*Beautiful*	21
Will	2.372.215	Work	56	*Woman*	759.817	*Friend*	38	***Different***	413.578	***Different***	20
Would	2.349.400	May*	38	*World*	732.511	*Night***	35	*American*	410.698	*Hot*	20

Fonte: elaboração dos autores.

**Thank com o mesmo número de ocorrências.*
***Morning com o mesmo número de ocorrências.*
Fonte em negrito ilustra palavras que aparecem tanto no COCA Corpus quanto no e-Tec Idiomas entre as 10 mais frequentes.

Considerando essa pequena amostra, verificamos que seis verbos, quatro substantivos e cinco adjetivos mais frequentes no material didático correspondem aos lemas mais frequentes da língua. Também é importante comentarmos sobre lemas, tais como *play, work, park, student, college, class, friend*, que parecem se circunscrever ao campo temático universitário, o qual faz parte da história que permeia as lições do Módulo 1 de Inglês do e-Tec Idiomas.

4.2 Relação *type/token*

No que diz respeito à relação *type/token*, ou seja, à quantidade de palavras diferentes (*types*) apresentadas aos estudantes e ao número total de dados (*tokens*), temos 2.618 *types* em 10.604 *tokens*, o que resulta em um TTR geral no material de 0,246. A seguir, Tabela 5, apresentamos os dados divididos por classe de palavra.

Tabela 5. TTR de verbos, substantivos e adjetivos.

Verbos			Substantivos			Adjetivos		
Types	Tokens	TTR	Types	Tokens	TTR	Types	Tokens	TTR
696	4.462	0,155	1.459	4.4663	0,312	463	1.479	0,313

Fonte: elaboração dos autores.

Como mencionado na seção de Metodologia, o resultado de TTR varia entre 0 e 1. Quanto mais próximo de 0, temos um contexto de maior repetição; quanto mais próximo de 1, maior é a diversidade lexical apresentada. Norberg e Nordlund (2018) apontam que não há um valor estabelecido sobre qual seria uma relação *type/token* ideal para materiais de aprendizagem de línguas, mas podemos traçar comparativos entre estudos empíricos já realizados com esse tipo de material.

Para avaliar a diversidade lexical em textos de diferentes tamanhos, Norberg e Nordlund (2018) utilizaram a Razão Tipo-*Token* Padronizada (*Standardized Type-Token Ratio* - STTR), que considera segmentos uniformes de palavras. Os livros didáticos analisados pelas pesquisadoras apresentaram STTRs que variaram entre 0,23 e 0,33. Na mesma linha, Criado e Sánchez (2012), ao analisarem dois livros didáticos de nível básico, apresentaram resultados de STTR geral de 0,30 e 0,33. No que tange, mais especificamente, às diferenças de TTR entre classes de palavras, os livros didáticos analisados por Norberg e Nordlund (2018) variaram entre 0,12 e 0,34 nos verbos; 0,20 e 0,47 nos substantivos e 0,13 e 0,38 nos adjetivos.

Em nosso material, os resultados de TTR de 0,15 para verbos e 0,31 para substantivos e adjetivos apontam mais oportunidades de repetição para a classe verbal que para as classes nominais. Faz-se importante mencionar novamente que, na análise de verbos, temos a seleção, tanto dos lexicais quanto dos auxiliares, considerando essa limitação de classificação no COCA *Corpus*. Caso estivéssemos analisando apenas verbos lexicais, o número de TTR certamente seria maior. Essa limitação da amostra tem relevância linguístico-analítica, em nosso entendimento, já que verbos auxiliares são interpretados como palavras gramaticais, com inventário reduzido e em geral muito frequentes nas línguas do mundo, ao passo que verbos de conteúdo são entidades de natureza lexical, limitadas pela demanda de informação expressa por seu conteúdo nocional (Aitchison, 1987; Cook, 2016).

A seguir, Tabela 6, com base na análise de Catalán e Francisco (2008), apresentamos os 50 *types* mais frequentes, entre verbos, substantivos e adjetivos, do COCA *Corpus* e do material didático do módulo 1 do e-Tec Idiomas, juntamente com o número de ocorrências.

Tabela 6. Os 50 *types* mais frequentes no COCA *Corpus* e no e-Tec Idiomas.

COCA Corpus						e-Tec Idiomas					
R	P	F	R	P	F	R	P	F	R	P	F
1	*is*	10.093.608	26	*see*	1.255.990	1	***is***	494	**26**	*morning*	35
2	*was*	6.848.519	27	*go*	1.235.814	2	***are***	328	**27**	*nice*	35
3	*'s [be]*	6.303.682	28	*going*	1.218.389	3	***do***	200	**28**	*friends*	34
4	*be*	5.046.995	29	*good*	1.111.721	4	***'s***	188	**29**	*night*	34
5	*have*	5.023.456	30	*way*	1.099.245	5	***have***	180	**30**	*work*	34
6	*are*	4.983.145	31	*want*	1.081.592	6	***can***	140	**31**	***were***	33
7	*do*	4.501.047	32	*got*	1.059.927	7	***'m***	125	**32**	***did***	32
8	*had*	2.723.573	33	*years*	1.032.285	8	**go**	79	**33**	***has***	32
9	*can*	2.515.641	34	*make*	1.023.483	9	**going**	70	**34**	*need*	32
10	*has*	2.443.889	35	*'ve*	1.015.706	10	**does**	63	**35**	*year*	32
11	*were*	2.378.665	36	*new*	1.000.338	11	***get***	57	**36**	*home*	31
12	*would*	2.349.400	37	*'ll*	992.384	12	*people*	55	**37**	*old*	30
13	*will*	2.154.429	38	*does*	964.997	13	***time***	54	**38**	*city*	29
14	*know*	2.110.629	39	*say*	955.726	14	*play*	52	**39**	***see***	29
15	*said*	2.039.196	40	*should*	920.908	15	***take***	51	**40**	*students*	29
16	*'re*	1.957.882	41	*come*	912.036	16	**was**	51	**41**	*winter*	28
17	*did*	1.889.734	42	*take*	858.510	17	***be***	49	**42**	*basketball*	26
18	*been*	1.874.157	43	*says*	837.850	18	***day***	42	**43**	*street*	26
19	*people*	1.782.910	44	*'d [have]*	784.592	19	***new***	40	**44**	thanks	25
20	*get*	1.743.855	45	*may*	757.742	20	***good***	38	**45**	***think***	25
21	*'m*	1.716.283	46	*let*	747.708	21	***may***	38	**46**	*clothes*	24
22	*time*	1.669.431	47	*man*	741.815	22	*park*	37	**47**	*house*	24
23	*could*	1.529.795	48	*life*	719.311	23	*am*	35	**48**	*library*	24
24	*think*	1.483.944	49	*day*	715.593	24	*college*	35	**49**	*meet*	24
25	*other*	1.278.825	50	*world*	714.815	25	*great*	35	**50**	***would****	24

Fonte: elaboração dos autores.

R = Ranking, P = Palavra, F = Frequência (número de ocorrências)
**Name, school, travel com o mesmo número de ocorrências.*
Fonte em negrito ilustra palavras que aparecem tanto no COCA Corpus quanto no e-Tec Idiomas.

Dos 50 *types* mais frequentes do COCA *Corpus*, 39 são verbos, e apenas 11 são não verbos: oito substantivos (*people, time, way, years, man, life, day, world*) e três adjetivos (*other, good, new*). Já em relação ao material analisado, temos 27 verbos e 23 não verbos: 18 substantivos (*people, time, day, park, college, morning, friends, night, year, home, city, students, winter, basketball, street, clothes, house, library*) e cinco adjetivos (*new, good, great, nice, old*). No geral, são 26 os *types* que coincidem com os 50 mais frequentes, tanto no COCA *Corpus* quanto no e-Tec Idiomas, dentre os quais apenas cinco são não verbos: *people, time, day, new, good*.

5 ANÁLISE QUALITATIVA

Em artigo intitulado "Vocabulary coverage in Spanish textbooks: How representative is it?", Davies e Timothy (2006) apontam que, comparativamente a um *corpus* de referência, palavras sub-representadas nos livros didáticos por eles analisados tendem a exprimir conceitos abstratos, tais como *momento, situação, necessidade, ocorrer, aparecer*, enquanto palavras

super-representadas são relativas a conceitos concretos. Os pesquisadores explicam que essa diferença ocorre porque o vocabulário dos livros didáticos costuma ser organizado em campos semânticos pré-definidos. Segundo eles, os autores de materiais didáticos agrupam vocabulário, textos e atividades a partir de temas como universidade, família, compras, casa, clima, meio ambiente, alimentação, saúde e lazer. Como os conceitos concretos se encaixam de forma mais direta nesses campos, acabam sendo super-representados, uma vez que os autores precisam ampliar o vocabulário ligado aos tópicos selecionados (Davies; Timothy, 2006).

Esse também parece ser o caso com o material didático aqui analisado. Palavras relativas a campos semânticos que visam satisfazer necessidades concretas estão entre os principais lemas na faixa de frequência acima de 5.050. Em relação aos verbos, alguns exemplos são *cycle, hike, jog, skate, ski, surf*, os quais fazem parte da seção que aborda atividades de lazer, assim como os verbos *mop (the floor), mow (the lawn), rake (the leaves)* são relativos a atividades cotidianas de cuidado com a casa.

No que diz respeito aos substantivos, exemplos abundam: membros da família (*stepdaughter, stepfather, nephew, niece*); características corporais (*slender, skinny, plump, overweight*); vestuário (*waistcoat, sweatshirt, sleepers, scarf, sandals, sweater*); utensílios e aparelhos domésticos (*cupboard, dishwasher, fireplace, heater, toaster, spoon*); e problemas de saúde (*chickenpox, backache, diarrhea, earache, faint, flu, insomnia, indigestion, pneumonia, rubella, stomachache*). Além disso, substantivos como *consulate, visa, passport, undergraduate, semester, seminar, pub*, também na faixa de frequência acima de 5.050, são relativos, mais especificamente, à temática abordada ao longo do Módulo 1 do material didático do e-Tec, que apresenta uma intercambista brasileira em um contexto universitário nos EUA. Já sobre os adjetivos nesta mesma faixa de frequência estão, principalmente, nacionalidades: *Argentinian, Brazilian, Australian, Danish, Finnish, Norwegian, Peruvian, Hungarian*.

Assim como os verbos, os substantivos e os adjetivos na faixa de frequência acima de 5.050 têm o intuito de preencher lacunas temáticas, fornecendo mais opções de vocabulário aos estudantes. É provável que isso esteja diretamente relacionado aos objetivos do nível A1 apresentados no QECR , o qual visa satisfazer necessidades imediatas e concretas: apresentar a si e aos outros, falar sobre aspectos pessoais, tais como o local onde vive, as pessoas que conhece, as coisas que têm. Obviamente, nem sempre as várias opções de vocabulário de áreas temáticas concretas coincidem com as palavras mais frequentes presentes nos grandes *corpora*.

Além de preencher lacunas temáticas, palavras na faixa de frequência acima de 5.050 também são expostas para preencher lacunas gramaticais. Exemplos disso estão na seção de *plural forms* (*cherry/cherries, tornado/tornadoes, cargo/cargoes, buffalo/buffalos/buffaloes*) e na seção de comparativos e superlativos (*windy/windier/windiest, breezy/breeziest, cloudy/cloudier/cloudiest, calmer/calmest, messier/messiest*).

Faz-se interessante apontar que, dentre os 1.103 *tokens* que pertencem à faixa de frequência acima de 5.050, 704 deles, ou seja, quase 64% são expostos na seção de explicação linguística, a qual é uma seção importante, tendo em vista que é dedicada à apresentação de vocabulário e gramática, dando destaque a determinados aspectos da língua. Como vimos, esse destaque, em alguns momentos, é direcionado a palavras de menor relevância em termos de frequência lexical.

6 CONSIDERAÇÕES FINAIS

O objetivo deste artigo foi analisar a frequência lexical de verbos, substantivos e adjetivos presentes nas apostilas de Inglês Módulo I do e-Tec Idiomas, comparativamente a um *corpus* representativo dessa língua, o COCA *Corpus*, tendo em vista a relevância de palavras de alta frequência no processo de aprendizagem de uma língua adicional.

No que diz respeito à distribuição de lemas por faixas de frequência, 85,59% dos verbos presentes no material didático do e-Tec Idiomas pertencem aos 1.000 lemas mais frequentes

da língua inglesa. Ao analisarmos substantivos e adjetivos nessa mesma faixa de frequência, esse número é menor: apenas 45,66% e 45,44%, respectivamente. No que diz respeito aos lemas na faixa intermediária de frequência (entre 1.001 e 5.050), temos 11,77% de verbos, 39,01% de substantivos e 36,31 de adjetivos. Em relação aos lemas que estão na faixa de frequência acima de 5.050, temos 2,64% de verbos, 15,33% de substantivos e 18,26% de adjetivos. Ao considerarmos a importância dos lemas na primeira faixa de frequência, vemos que o resultado dos verbos é mais positivo que o resultado das classes nominais. Uma maior porcentagem de verbos na primeira faixa de frequência comparativamente a substantivos e adjetivos também foi o resultado de Norberg e Nordlund (2018).

No que concerne à relação *type/token*, o resultado geral, incluindo verbos, substantivos e adjetivos, foi de 0,246, o qual parece adequado para um material didático de nível básico. Ao analisarmos separadamente as classes de palavras, temos resultados de TTR de 0,15 para verbos e 0,31 para substantivos e adjetivos. Com isso em vista, há mais oportunidades de repetição para a classe verbal que para as classes nominais. No entanto, faz-se importante fazermos, novamente, uma ressalva em relação aos resultados dos verbos pelo fato de o COCA *Corpus* não separar verbos auxiliares dos lexicais.

Nossa análise qualitativa demonstrou que a maioria das palavras na faixa de frequência acima de 5.050 preenchem lacunas temáticas e gramaticais, com o intuito tanto de satisfazer os objetivos relativos ao nível A1 do QERL (CEFR, Council of Europe, 2001) quanto de contemplar a temática relativa a um intercâmbio universitário, o qual faz parte da história que permeia as lições do Módulo 1 de Inglês do e-Tec Idiomas. Assim como apontam Norberg e Nordlund (2018), textos compostos por palavras apenas de alta frequência seriam, muito provavelmente, desinteressantes e inadequados, pois alguns temas exigem vocabulário menos frequente: a questão, portanto, é encontrar um equilíbrio entre palavras de alta e baixa frequência que favoreça o aprendizado.

Por fim, faz-se importante ponderarmos sobre as especificidades de cada classe de palavra. Substantivos e adjetivos costumam apresentar maior variação, a depender do contexto no qual estão inseridos: são palavras mais sensíveis à temática e à situação de uso. Diferentemente, verbos de alta frequência, por exemplo, são utilizados em uma ampla variedade de contextos: essa flexibilidade faz com que convirjam mais facilmente com a distribuição lexical dos *corpora* gerais. Nesse sentido, mais pesquisas de análise lexical em livros didáticos, análogas a essa, são necessárias para traçarmos paralelos e refletirmos de forma mais acurada sobre os resultados aqui apresentados.

AGRADECIMENTOS

Agradecemos o suporte oferecido pelas instituições de ensino envolvidas e a bolsa de Iniciação Científica concedida pela Pró-reitoria de Pesquisa do Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense, Câmpus Novo Hamburgo-RS, Edital 09/2022.

REFERÊNCIAS

AITCHISON, J. *Words in the mind*: an introduction to the mental lexicon. Oxford: Blackwell, 1987.

CATALÁN, R. M. J.; FRANCISCO, R. Vocabulary input in EFL textbooks: frequency and distribution of words. *International Journal of Applied Linguistics*, *[S. l.]*, v. 18, n. 1, p. 49-70, 2008.

COOK, V. *Second Language Learning and Language Teaching.* 5th ed. New York: Routledge, 2016.

COUNCIL OF EUROPE. *Common European Framework of Reference for Language Learning, Teaching and Assessment*. Cambridge: Cambridge University Press, 2001.

CRIADO, R.; SÁNCHEZ, A. Lexical frequency, textbooks and learning from a cognitive perspective: a *corpus*-based sample analysis of ELT materials. *Revista Española de Lingüística Aplicada*, *[S. l.]*, v. 25, n. 1, p. 77-94, 2012.

DAVIES, M.; FACE, T. L. Vocabulary coverage in Spanish textbooks: How representative is it? *In*: SAGARRA, N.; TORIBIO, A. J. (ed.). *Selected Proceedings of the 9th Hispanic Linguistics Symposium*. Somerville, MA: Cascadilla Proceedings Project, 2006. p. 132-143.

E-TEC IDIOMAS SEM FRONTEIRAS. Brasília, DF: Instituto Federal de Educação, Ciência e Tecnologia, [s.d.]. Disponível em: http://cpte.ifsul.edu.br/docs/e-tec_idiomas_semfronteiras.pdf. Acesso em: 18 jan. 2025.

LIMA, J. C. *et al*. *English*: Module 1, Book 3. Pelotas, RS: IFSul, 2015.

MOREIRA, H. B. *et al*. *English:* Module 1, Book 1. Pelotas, RS: IFSul, 2015.

NATION, I. S. P. How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *[S. l.]*, v. 63, n. 1, p. 59-82, 2006.

NATION, I. S. P. Vocabulary research into practice. *Language Teaching*, *[S. l.]*, v. 44, n. 4, p. 529-539, 2011.

NORBERG, C.; NORDLUND, M. A *corpus*-based study of lexis in L2 English textbooks. *Journal of Language Teaching and Research*, *[S. l.]*, v. 9, n. 3, p. 463-473, May 2018.

PEREIRA, A. N. *et al*. *English*: Module 1, Book 2. Pelotas, RS: IFSul, 2015.

SAKATA, N. Profiling vocabulary for proficiency development: effects of input and general frequencies on L2 learning. *System*, *[S. l.]*, v. 87, p. 1-12, 2019.

Contribuição dos autores

Este artigo reporta resultados de estudo realizado em estágio pós-doutoral pela primeira autora na Universidade Federal do Rio Grande do Sul (UFRGS), entre os anos 2023 e 2024, sob a supervisão do segundo autor e em colaboração com o terceiro autor.