ANA LÚCIA SURERUS PITANGUY MARQUES

Science and Geography lexicons in English for learners in "Ensino Fundamental 1": a corpus-based investigation

Universidade Federal de Minas Gerais – UFMG
Belo Horizonte
2024

UNIVERSIDADE FEDERAL DE MINAS GERAIS – UFMG
FACULDADE DE LETRAS – FALE
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS LINGUÍSTICOS

ANA LÚCIA SURERUS PITANGUY MARQUES

Science and Geography lexicons in English for learners in **"Ensino Fundamental 1": A corpus-based investigation**

Tese apresentada ao Programa de Pós-Graduação em Estudos Linguísticos da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do título de doutor em Linguística Aplicada.

Área de concentração: Linguística de Corpus Aplicada.

Linha de Pesquisa: Ensino / Aprendizagem de Línguas Estrangeiras – 3A

Orientadora: Profa. Dra. Deise Prina Dutra.

Belo Horizonte
2024

M357s

Marques, Ana Lúcia Surerus Pitanguy.
Science and Geography lexicons in English for learners in "Ensino Fundamental 1" [manuscrito] : a corpus-based investigation / Ana Lúcia Surerus Pitanguy Marques. – 2024.
1 recurso online (219 f.: il., grafs., tabs., color, p&b.) : pdf.

Orientadora: Deise Prina Dutra.

Área de concentração: Linguística de Corpus Aplicada.

Linha de pesquisa: Ensino/Aprendizagem de Línguas Estrangeiras - 3A

Tese (doutorado) – Universidade Federal de Minas Gerais,

Faculdade de Letras.

Bibliografia: f. 127-139.

Apêndices: f. 140-219.

1. Linguística de corpus – Teses. 2. Língua inglesa – Estudo e ensino – Falantes estrangeiros – Teses. 3. Língua portuguesa (Ensino fundamental) – Estudo e ensino – Teses. I. Dutra, Deise Prina. II. Universidade Federal de Minas Gerais. Faculdade de Letras. III. Título.

CDD: 420.7

Ficha catalográfica elaborada pela bibliotecária Priscila Oliveira da Mata – CRB/6-2706
Biblioteca Professor Rubens Costa Romanelli - FALE/UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS
FACULDADE DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS LINGUÍSTICOS

FOLHA DE APROVAÇÃO

Science and Geography lexicons in English for learners in "Ensino Fundamental 1": a corpus-based investigation

ANA LÚCIA SURERUS PITANGUY MARQUES

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTUDOS LINGUÍSTICOS, como requisito para obtenção do grau de Doutor em ESTUDOS LINGUÍSTICOS, área de concentração LINGUÍSTICA APLICADA, linha de pesquisa Ensino/Aprendizagem de Línguas Estrangeiras.

Aprovada em 15 de fevereiro de 2024, pela banca constituída pelos membros:

Prof(a). Deise Prina Dutra - Orientadora

UFMG

Prof(a). Paula Tavares Pinto

UNESP

Prof(a). Bárbara Malveira Orfanó

UFMG

Prof(a). Shirlene Bemfica de Oliveira

IFMG

Prof(a). Lucia de Almeida Ferrari

UFMG

Belo Horizonte, 15 de fevereiro de 2024.

Documento assinado eletronicamente por **Deise Prina Dutra**, **Professora do Magistério Superior**, em 16/02/2024, às 14:27, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Barbara Malveira Orfano**, **Professora do Magistério Superior**, em 16/02/2024, às 14:36, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Lucia de Almeida Ferrari**, **Professora do Magistério Superior**, em 16/02/2024, às 15:00, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Shirlene Bemfica de Oliveira**, **Usuário Externo**, em 19/02/2024, às 15:09, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Paula Tavares Pinto**, **Usuária Externa**, em 19/02/2024, às 21:26, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2911726** e o código CRC **73A87D1B**.

For Daniel, Thomas, Rafael, Andrew, and Catarina, my grandchildren, a glimpse into my journey in a new world: the uncharted territory of Applied Corpus Linguistics.

Accomplishments in life depend mostly on what we believe in and fight for!

Acknowledgements

I wish to express my heartfelt thanks to my inspiring supervisor, Dr. Deise Prina Dutra, for having taken a leap of faith when she invited me to join her group of postgraduate students in 2019. Her warm welcome opened a new perspective in my life, offering a unique opportunity to my return to university studies. Soon after, the admission process was set in motion, and fortunately, it unfolded much more positively than I had expected. At that time, I was totally unaware that an imminent danger was approaching fast: the dawning of a pandemic that would change our lives and lifestyles forever. Nevertheless, the commitment to the studies kept me busy, helping me go through that somber period unharmed.

I am especially thankful to Clara, the teacher, who agreed unconditionally to embrace my investigation and implement it in her six groups of Elementary students. I am very grateful for her diligence in making it happen according to the guidelines, certainly a decisive factor in the positive outcomes of the study.

My gratitude to Dr. Ricardo de Souza for his thorough description of my work, approving it without restrictions, which allowed me to advance towards my doctoral degree.

I would like to express my gratitude to Dr. Paula Tavares Pinto, Dr. Bárbara Orfanó, Dr. Lucia Ferrari and Dr. Shirlene Bemfica for their time, effort, and expertise. Your reviews and insightful feedback have significantly contributed to refining of my thesis. I am particularly grateful for your constructive criticism, which has not only improved my work but also my skills as a researcher.

A very special thank you to Dr. Bárbara Orfanó, Dr. Carla Coscarelli, and Dr. Luciane Ferreira for your inspiring lessons, which helped me thread on *roads not taken* before. They led me into the world of data-driven learning, ICT literacy, and the world of Portuguese as a welcoming language. I have greatly enjoyed and benefited from your lessons.

I would also like to thank my husband for his unwavering support through this process. Homebound for three years, I was able to spend my time meaningfully carrying out an investigation on the web and in the books I had acquired. The research work enabled me to attend courses and meetings online, present papers at conferences in faraway places, and meet scholars from whom I learned so much.

My unconditional love to my children Erika and Henrique for their spreading enthusiasm regarding my study along these past four years.

Abstract

Corpora-based studies have influenced language education, especially for adults, since the 1980s with dictionaries and grammar books (Sinclair, 1987, 1990; Biber *et al*., 1999). However, there is still a paucity of research on how corpus-based materials can support primary school learners of English as an additional language. With the growth of bilingual programs (Portuguese-English) in our context (Oliveira; Höfling, 2021), there has been an increasing demand for materials based on real language use in the past years. In an attempt to address the demand, this investigation is about using specially compiled pedagogic corpora that can accelerate learners' exposure to English in subjects like Science and Geography. The outcomes of the study can reveal benefits of integrating corpus-informed pedagogy in the early levels of *Ensino Fundamental I* (Elementary school) in Brazil. Despite the fact that there are a few studies based on corpus for this educational level (Hirata, 2020; Crosthwaite; Stell, 2020), mainly using the data-driven learning (DDL) approach (Johns, 1991), none had the same learners' academic level, nor a similar linguistic focus or aim. To the best of our knowledge, there has not been any research on the benefits of the use of pedagogic corpora which address the linguistic needs of 4th-6th primary school students. To compile the corpora, first, topics in the domains of Science and Geography were chosen from the Brazilian National Common Curriculum (BNCC) guidelines for teachers. Second, texts on the chosen topics were selected from textbooks and websites. Third, texts were level checked (A1-A2), according to the Common European Framework of Reference for Languages (CEFR). A balanced number of texts composed the corpora, considering school year and source (textbook or web), yielding two corpora: (a) Science with 437 texts and (b) Geography with 458 texts. Two topics were selected for the study, one from each corpus: *Neighborhood* from Geography and *Animals* from Science. Fifth, a lexico-grammatical analysis of the two subsets of corpora was done by the concordancer (#LancsBox), revealing the most frequent content words and lexical bundles which compose our year 4-6 grade *Neighborhood* and *Animals* Word List. And finally, the teacher and the researcher identified the vocabulary which would be most appropriate and meaningful for young learners 9-12 years old, and activities were specially designed. The DDL approach was used to implement all the activities with concordance lines and key words in context (KWICs) in the classrooms. The resulting outcomes of learners' production have yielded positive insights into the feasibility of working with authentic language in the initial levels of the elementary school. Indeed, the subject-focused level-appropriate lexicon compiled

can complement learners' exposure to English and support teachers' work. The concordance lines added variety and change to the class routine motivating learners to walk the extra mile to play with them and understand the meaning of the key words. It is fair to say that this repeated *condensed exposure* (Gabrielatos, 2005) can contribute to an early L2 vocabulary expansion and learning and heighten awareness of language patterns (Granger, 1998).

Keywords: concordancer; English textbooks; Geography and Science; pedagogic corpora; KWIC; concordance lines.

Resumo

Os estudos baseados em *corpora* têm influenciado a aprendizagem de línguas, em especial, por adultos, desde a década de 1980, com a publicação de dicionários e livros de gramática (Sinclair, 1987, 1990; Biber *et al*. 1999). Entretanto, ainda existe escassez de pesquisas sobre como os materiais criados a partir de *corpora* podem ser benéficos aos alunos de inglês como língua adicional no ensino Fundamental I. Mesmo assim, com o aumento dos programas bilíngues (Português-inglês) no nosso contexto (Oliveira; Höfling, 2021), tem ocorrido o crescimento na demanda de materiais baseados em língua autêntica nos últimos anos. Em uma tentativa de responder a essa demanda, esta investigação trata do uso de corpora pedagógica, especialmente compilada, que poderá acelerar a exposição ao inglês de jovens aprendizes em disciplinas como Ciências e Geografia. Os resultados do estudo indicam benefícios de se integrar um modelo de pedagogia fundamentado em corpus aos anos iniciais do ensino Fundamental I no Brasil. Apesar de haver alguns estudos com o uso de corpora no nível secundário (Hirata, 2020; Crosthwaite; Stell, 2020), a maioria usando a abordagem *ensino baseado em dados de corpora* (*data-driven learning*) (Johns, 1991), nenhum foi feito com o mesmo nível escolar dos participantes, com o mesmo objetivo ou foco linguístico. Ao que parece, até esta data, nenhuma pesquisa foi feita para que fossem identificados os benefícios do uso de *corpora* pedagógicas que atendam às necessidades linguísticas de alunos da 4ª., 5ª. e 6ª. séries da escola primária. Para compilar os *corpora* pedagógicos, primeiro foram escolhidos tópicos de Ciências e Geografia, obtidos nas orientações da Base Nacional Comum Curricular (BNCC). Depois, textos e vídeos nos temas selecionados foram extraídos de livros-texto e de *websites*. Como terceira etapa, o material passou pela avaliação dos níveis linguísticos A1–A2 do *Common European Framework of Reference for Languages* (CEFR), indicado para os níveis iniciais do Ensino Fundamental I. Na etapa seguinte, os dois *corpora* foram compilados com um número balanceado de textos, agrupados por série escolar e fonte do material (*web* ou livros), gerando (a) Ciência com 437 textos e (b) Geografia com 458 textos. Para esta investigação foram escolhidos dois tópicos: Vizinhança (*Neighborhood*), de Geografia, e Animais (*Animals*), de Ciências. A quinta etapa foi feita pelo concordanceador *#LancsBox*, que realizou a análise léxico-gramatical nos dois subgrupos dos *corpora* e identificou os vocábulos e os grupos de palavras mais frequentes em cada *subcorpus*. E, por último, esta pesquisadora e a professora participante identificaram o vocabulário que seria mais apropriado e significativo para os alunos de 9 a 12 anos, para que fossem criadas as atividades especiais. A abordagem

em sala de aula foi o DDL usado na implementação das atividades com as linhas de concordância e palavras-chave (KWIC). Os resultados da produção linguística dos alunos revelaram aspectos positivos para a continuação do trabalho com língua autêntica em níveis iniciais do Ensino Fundamental I. Um léxico compilado com conteúdo adequado aos temas e séries dos alunos pode, sim, complementar a exposição desses alunos ao inglês e ao trabalho dos professores. As linhas de concordância adicionaram variedade e mudança na rotina das aulas durante o estudo, motivando os alunos a se esforçarem para compreender o significado das palavras-chave ao manuseá-las. Parece ser razoável afirmar que esta exposição condensada (Gabrielatos, 2005) pode contribuir para uma expansão e aprendizagem precoce do vocabulário alvo e para o aumento da percepção das estruturas da língua adicional (Granger, 1998).

Palavras-chave: concordanceador; livros-texto em inglês; Geografia e Ciências; *corpora* pedagógicos; KWIC; linhas de concordância.

LIST OF FIGURES

LIST OF TABLES

LIST OF CHARTS

TABLE OF CONTENTS

Chapter 1 – Introduction

1.1 Context

The world has been undergoing fast and irreversible changes in the last decades and formal education has been called in to adjust its priorities to the demands of the new generation (*O Estado de São Paulo*, June 2, 2019). Different media have been emphasizing the fact that we are now dealing with a generation of students who learn in a much different way than 10 years ago. Articles, documentaries, and those investigating our educational system in general are urging educators to understand learners' new dynamics and organize learning so that it benefits everyone. The most recent generation of young learners, currently in elementary school, is certainly one that challenges the learning / teaching boundaries of the past even further while demanding teachers' mentoring in new ways in the classroom. Formal education has to respond fast if it aims to prepare the country's youngsters to fit in successfully in the new marketplace format after the school years.

To this date, teachers' roles have been multi-faceted, trying to provide learners not only with the subject-matter contents of their lessons but also guiding them towards meeting learning goals. However, most of those roles are now being disputed as technology and portable devices, available to a large portion of the population, offer learners instantly the information required for their day-to-day lives. It is the turning point of the source of knowledge: the tools available today can supply the present generation of learners with the right answers at the tip of their fingers. In Clemesha and Liberali's (2020, p. 1) words, "we believe that education should provide expansive new modes of effective participation in society to offer students the chance to increasingly develop forms of insertion in the world and means to transform their mobility." Contemporary education should enable learners to identify the schools' goals to engage more effectively with them, to be motivated to be in the school and be ready to learn.

It is of utmost relevance to begin describing this study by mentioning that back in 1996, the Ministry of Education (*Ministério da Educação e Cultura* – MEC) had already taken steps towards setting further parameters for education by launching what is called *Base Nacional Comum Curricular* (BNCC).[1] The first BNCC was a 600-page document which tried to unify the school system by defining its academic contents, aligning the regulations for proper

[1] Most recent version of BNCC was validated on December 14, 2018. Available at: http://basenacionalcomum.mec.gov.br/implementacao/.

implementation of related policies in schools throughout the country. Among other issues, it contemplates all significant areas of knowledge required for the *Fundamental* levels of all schools. Under these new circumstances, the schools had to accommodate the innovations and teachers have been witnessing curricula adjustments. These have been made fast in both private and public schools to meet the latest requirements, with one eye on the learners and another in the future. What policy makers need to realize is that the generation which is going to be responsible for the future of the country is being prepared now.

The key role for educators is to prepare students for life, but at a time of great change and uncertainty, are teachers prepared to teach what is necessary? It is common knowledge, and current statistics show, that English as an additional language (EAL) and digital literacy, knowledge of basic technological tools, are the new assets every young adult needs to have when leaving school to join the workforce. To illustrate the point, words in English like *lockdown*, *home office*, *co-working*, *shared workplace*, *co-living*, *co-housing, coaching* and even the more recent *home-schooling,* are just a few of those which are now part of everyone's new lifestyle. Many schools are trying hard to respond to the new reality and taking as fast an action as they are able to in order to address their stakeholders' demands: better-informed parents want their children to be well-prepared now for their future professional lives. They rely on formal education because most of them believe the younger their children are exposed to English, the better their English competence will be in the future.

It resonates with Read's words when addressing an IATEFL[2] conference (2000, p.33) on 'Young Learners & Teenagers', who declared twenty-three years ago that

> the last 30 years [had] seen a continuous driving down of the age of introducing English as a school subject. […] the main driver behind this pervasive global trend has been the potential political and economic benefits perceived by governments and a public keen to give their children an educational edge for the future.

She was referring to the noticeable phenomenon taking place elsewhere in the world, mainly in Asia, at that time. The same phenomenon has been observed to a greater or lesser extent in Brazil in more than a decade, with private regular schools gradually preparing themselves to become bilinguals mainly in the capital cities. One such example is the 'Multicultural Breakfasts and Translanguaging Kids' Project (Clemesha, Liberali, 2020), currently taking place at a school in São José dos Campos, São Paulo, which aims at having a broader

[2] International Association of Teachers of English as a Foreign Language.

participation of their learners in society. It follows an inquiry-based curriculum, which is organized through transdisciplinary projects in the primary years in an attempt to create a multilingual curriculum (Liberali, 2019).

Nonetheless, Oliveira and Höfling (2021, p. 6) claim that

> Brazil currently does not have an approved education law or policies regarding (Portuguese-English) bilingual programs. [Only] in June 2020, the *Conselho Nacional de Educação* (CNE) [National Council of Education] and the *Conselho de Educação Básica* (CEB) [Council for Basic Education] launched a white paper presenting a proposal for curricular guidelines for bilingual education (Brasil, 2020).

This means there is still a long way to go before analyses and thorough examination of educational aspects by those involved can afford a final version of the document to be released, making it possible for bilingualism to be considered an official option in primary schools.

The second asset our learners need to have is literacy in information and communication technology (ICT) as stated in the Brazilian National Syllabus Core (Brasil, 2018, p. 9),

> 5. Understand, use and create digital information and communication technology,[3] in a meaningful, reflexive and ethical way, in the various social practices (including the school ones) to communicate, access and share information, create knowledge, solve problems and take agency of one's own personal and collective life.[4]

According to BNCC, learners also need to have general ICT skills to look for information on the web. They need to be acquainted with the web search engines appropriately to navigate and learn how to find what they need safely and effectively. Ribeiro (2020) mentions other factors of great relevance to be discussed by public policies and implemented such as the availability of hardware in public and private schools (Mendes; Finardi, 2020), the dearth of in-service training on the use of technology with a focus on educational development of learners and also misinformation and prejudice against the use of computers in schools.

A better prepared and skilled young adult would certainly stand a better chance to succeed and be prepared to seek further opportunities in the future. In times of great cultural awareness and the need for social equity and inclusion in regular schools, English and digital

---

[3] ICT competence.

[4] Authors' translation for: 5. Compreender, utilizar e criar tecnologias digitais de informação e comunicação de forma crítica, significativa, reflexiva e ética nas diversas práticas sociais (incluindo as escolares) para se comunicar, acessar e disseminar informações, produzir conhecimentos, resolver problemas e exercer protagonismo e autoria na vida pessoal e coletiva (Brasil, 2018, p. 9).

literacy can be the key elements to help reduce the educational differences our youngsters experience while attending the primary school years. Learning resources and activities, among other aims, should "open up learning to new, real-world contexts, which involve learners themselves in hands-on activities, scientific investigation or complex problem solving, or in other ways, increase learners' active involvement in complex subject matters" (Redecker, 2017, p. 22). Instead of the present situation in many schools, young learners should be exposed to motivating content and learn to identify their preferences to choose their future pathways.

1.2 Justification and description of the problem

To empower young learners with digital skills, one avenue is the inclusion of instruction on the use of technology and its tools more regularly in the elementary schools curricula. Another is teachers supplementing the curricula with an additional area of interest during specific periods of the school calendar. As for the acceleration of young learners' English competence, more exposure to the language during the initial years of primary school can be beneficial and effective in the learning process. Read (2003, p. 6) says that what counts is not an *optimal age*, but *optimal teaching-learning conditions* emphasizing that

> primary schools generally provide an ideal context for a whole learning experience appropriately structured to meet children's needs. Through 'learning by doing', language competence can be built up gradually and naturally and provide the basis for more abstract, formal learning in secondary school.

She claims schools should implement "coherent primary and secondary policies to provide for progression and continuity throughout the school years, […] curricula should build on what children know by the end of primary school rather than require them to start again" (Read, 2003, p. 6).

In the talk "Using Corpora in the Language Classroom – The New School"[5] (2012), while addressing an audience of teachers and teacher trainers in New York, USA, Reppen tried to demystify the use of corpora in classrooms by showing it could empower teachers instead. Corpora, "a principled collection of texts available for qualitative and quantitative analysis" (Biber *et al.,* 1998) are stored in computers which allow for a very large amount of texts to be analyzed by software and inform researchers. The large-scale "corpora databases" can aid teachers and English Language Teaching (ELT) materials writers more effectively to address

[5] See: https://www.youtube.com/watch?v=Qf46lOnMCfs&t=967s.

the difficulties learners have to master a language's vocabulary. According to Reppen (2010, p. 10), the "different genres of authentic language [produced by corpora databases] have enabled professionals to design activities and materials that reflect authentic language use" as corpus-informed lexicography can bring natural language into the classroom in a way that it can involve learners through hands-on tasks.

It made me reflect and start speculating about a way to accelerate the use of English in other subjects other than General English, in an attempt to integrate corpus-informed pedagogy in our *Fundamental 1*[6] schools. If teachers of additional languages (ALs) could resort to topicalized corpora, i. e., corpora in different subjects in English (L2) to devise activities for cross-curricular projects, they would be able to increase young learners' exposure to L2 vocabulary in their early school years. As Webb and Chang (2012, p. 276) have put it, such exposure could trigger deliberate learning, for example, by working with "vocabulary through the completion of exercises and tasks where the primary aim of the activity is to learn target words, also known as intentional or explicit learning."

To support the view on a comprehensive early start in L2 acquisition, Pinter (2012) postulates that children from 8 years old go through an increase in their L1[7] competence and become more aware of how language works, which are favorable characteristics for the introduction of an additional language in their regular studies. However, available literature suggests that although that is positive, youngsters would only benefit from an early exposure to language given that some conditions are met. According to Marinova-Todd *et al*. (2000, p. 28),

> only if teachers are themselves […] well-trained in the needs of younger learners; if the learning opportunities are built upon with consistent, well-planned, ongoing instruction in the higher grades; and if the learners are given some opportunities for authentic communicative experiences in the target language.

To embrace and address the above regarding the introduction of English in primary schools, this investigation's outcomes, the subsets of corpora, will need to be made available to teachers. They will have to decide whether the intended audience of young learners will use the specific vocabulary to produce language in writing and speaking (active / productive) or will use it just for recognition and comprehension (passive / receptive) (Melka, 1997; Nation; Waring, 1997). Generally speaking, the term *young learner* is used to refer to children from

[6] Fundamental I (Brazil) and Elementary (US) years are equivalent and will be used interchangeably.
[7] In our case, Portuguese is the L1.

their first year of formal schooling, usually somewhere between 5-7 years old, to when they are 11-12 years old (Read, 2011)[8]. In sum, the investigation's findings should inform and meet L2 teachers' demands so they are able to use the corpora-informed material in the classroom.

For the pedagogical purposes of this study, I am going to use the terms vocabulary and lexis interchangeably to refer to individual words, collocations, idioms and fixed and semi-fixed expressions because vocabulary learning frequently involves learning "chunks"[9] (Lewis, 1993), or n-grams that are longer than individual words. The n-grams are groups of words which, in Sinclair's observations (2004, p. 29) on the *Idiom Principle* and the *phraseology tendency*, "do not appear in isolation but go together and make meanings by their combinations, such as collocations and other features of idiomaticity" (Sinclair, 2004, p. 9).

A corpus-based investigation should address those specificities by providing descriptions of actual language use, and the examples should be chosen to suit the interests of learners and be most relevant to their needs. McCarten (2007, p.26) emphasizes the need to "[start] with the most frequent, useful, and learnable vocabulary, returning later to more difficult vocabulary" and less frequent uses of previously learned items. McCarthy (2004) complements by saying that learning words with their most frequent collocations is a good learning habit that can start right from the lowest levels and recommends that teachers should mediate "data" to find the clearest and best examples to use from the corpus. Reppen (2010, p. 27) complements those remarks by saying "the investigation […] of co-occurrences of lexical patterns (n-grams) in topicalized lexicon can play a pivotal role in the acquisition of L2," provided such corpora is large enough to allow for all senses of a word to be represented, and allow learners to possibly guess its meaning when encountering it for the first time.

Therefore, in an attempt to contemplate the above issues, the study to be described in the next chapters investigates naturally-occurring existing lexicography in authentic L2 elementary school textbooks as well as texts, articles and website videos. They were compiled as a subject-focused level-appropriate lexicon to complement the L2 input learners are exposed to in *Ensino Fundamental 1* – grade 4 (9 years old), grade 5 (10-11 years old) and grade 6 (12 years old). Ultimately, the resulting core-aligned specialized[10] corpora (Aston, 1997) can advise teachers who will be able to resort to the information to design familiar activity types with authentic language (McCarthy, 2004). The examples of activities and procedures in the

---

[8] Available at: https://carolread.wordpress.com/. Section: "Y is for Young Learners." Accessed on: November 10, 2022.

[9] One possible definition of 'chunk' is "an all-purpose word that embraces any formulaic sequence, lexical/phrasal expression or multi-word item," mentioned by Lewis (1993) in his book The Lexical Approach.

[10] The adjectives topicalized and specialized corpus / corpora will be used interchangeably in this study.

classroom included in Chapter 3 should further motivate teachers to add variety and raise learners' interest in learning another language.

1.3 Aims of the study

The aim of this study is to test / show the feasibility of adopting vocabulary corpus-based activities in EAL class. In an attempt to show their effectiveness, pedagogic corpora was compiled to generate vocabulary for the tasks which will have a more focused approach on the target language when lessons are delivered.

This study has searched, selected, compiled, examined, classified, categorized, and built level-appropriate corpora-based lexicography in English. It has been designed to provide samples of tasks in specific topics selected from two *Areas of Knowledge*: Earth and Space Science and Life Science[11] proposed by the 2018 BNCC.[12] The compiled pedagogic corpora (Willis, 1998) with Science and Geography texts have allowed the researcher to identify the most frequent keywords and phrases and analyze linguistics features in context for the 4th to 6th school grades. It was done to meet the overriding aim of this corpus-based study which is to start exposing young learners to English at a much younger age than it takes place today. In order to achieve that, my main aims were:

1) identify the top most frequent individual content words and lexical bundles (Biber *et al.*, 1998) worked with in 4th to 6th grades in both corpus for elementary school on Geography (COREL-GEO) and corpus for elementary school on Science (COREL-SCI);
2) propose samples of activities designed with corpus-informed language that could serve as a model for primary school teachers who want to boost their students' English learning in a contextualized format; and
3) point out the extent to which the activities proposed might have led to vocabulary learning in the different grades.

In sum, this study has identified and compiled different kinds of content words such as nouns, verbs, adverbs, adjectives and chunks, lexical bundles, in specialized subsets of corpora, that are relevant in English to learners in 4th to 6th grades. The resulting high-frequency L2 lexicon should make it possible for teachers to gain insight into the authentic language learners will be exposed to. The possible availability of the findings – dedicated corpora in an online

---

[11] In Portuguese in the original guidelines: **Ciências da Natureza,** 1) Vida e Evolução e 3) Terra e Universo. **Ciências Humanas**, 1) O sujeito e o nosso lugar no Mundo e 5) Natureza, Ambiente e Qualidade de Vida.
[12]The original source is in Portuguese and main areas have been translated to English to contextualize the choices made by the researcher. Available at: https://novaescola.org.br/conteudo/12720/bncc-baixe-em-pdf-o-e-book-de-competencias-gerais.

platform – will be able to empower L2 teachers to generate the vocabulary and design tasks and / or implement cross-curricular projects in the future.

1.4 Research questions

To achieve the above mentioned aims, some questions were formulated to guide the research into the search for answers:

i) Which are the top most frequent topic-related L2 content words and 3- and 4-grams in COREL-GEO for 4th to 6th grades?

ii) Which are the top most frequent topic-related L2 content words and 3- and 4-grams in COREL-SCI for 4th to 6th grades?

iii) Can activities implemented with the data-driven learning (DDL) approach expand learners' topicalized vocabulary and possibly boost their progress in English?

iv) Are the results significantly different from one grade to the others when the same tasks are worked with in the classrooms?

Chapter 2 Theoretical Framework

2.1 Learning another language

How much vocabulary should L2 students learn per year? In past decades, L2 beginners would only be exposed to a basic vocabulary of two to three hundred words in most general English classrooms. Based on a previous study by Webb and Nation (2017, p. 232) suggest

> that in a principled vocabulary learning programme in [an English as a Foreign Language] (EFL)[13] context, learning 400 word families per year may be a realistic goal for all learners, [to enable them to have] a relatively comprehensive knowledge of these words through repeated encounters in spoken and written discourse, as well as frequent opportunities to use them.

According to one estimate, about 30% of research on vocabulary in the last 100 years has been carried out since 2001 (Nation, 2013). The findings yielded by this whole body of more recent research, in Victoria University of Wellington, Temple University Japan, Koran Women's Junior College, the University of Western Ontario (Nation, 2017) and Taiwan (Nation, 1983, 1990; Schmitt; Schmitt; Clapham, 2001), to name just a few, show that beginner learners are estimated to learn about 500 words on average in the primary years of school in an EFL context[14], a figure that allegedly does not suffice as it still prevents learners from establishing patterns of word uses (McCarthy, 2004). Although investigations differ in how they measure and assess L2 vocabulary growth, one of the starting points can be an analysis of the results of the mother tongue (L1) acquisition by learners: they may know at least 3-4,000 *word families* before they can read, at around five or six years of age.

By *word families* it is understood that they consist of a headword, its inflections and its derivations and because they include several *lemmas*, count headwords and inflections as the same item and derivations as separate items, the results of range, frequency and dispersion analyses are underestimated (Nation, 2006). The author claims that many studies show evidence that "the high-frequency and wide-range words are generally learned before lower-frequency and narrower-range words" (Nation, 2006, p. 63). According to him, it is assumed that both native- and non-native-speaking learners acquire vocabulary largely in the order of its range

---

[13] EFL is currently widely referred to as English as an Additional Language (EAL). In this study I will quote it as it is mentioned in the literature, even though I am referring to English as an additional language throughout the investigation.

[14] "The term 'English as a Foreign Language' (EFL) applies to learners who are living in a country where English is not the first or significant language; examples include Japan, Brazil, and France" (Nation, 2017. p. 155).

and frequency (West, 1953; Nation, 1990, 2001, 2013a). Nation and Waring (1997, p. 11) claim that "the learner needs to know the 3,000 or so *high-frequency* words of the language. These are an immediate high priority and there is little sense in focusing on other vocabulary until these are well learned." Webb and Nation (2017, p. 29) state further that

> [w]e typically learn high-frequency words in our first language incidentally, as we encounter them repeatedly in speech and writing. However, when we learn another language, we may need to deliberately learn most of these words. High frequency words have the greatest value for language learning, so they deserve attention in the classroom. […]

These 3,000 are the words encountered more regularly in all forms of speech or writing. To illustrate this argument, Webb and Nation (2017, p. 5) refer to "the books we read as children [which] are typically designed to promote vocabulary learning [and] that their pictures are provided to illustrate the meaning of key content words." Those nouns, adjectives, verbs, and adverbs are relevant because they convey meaning. Nation concludes that the challenge for beginning learners and readers is reaching to the threshold where they can learn from context. Additionally, he emphasizes that achieving the ability to read authentic L2 texts requires a larger vocabulary size.

One of the studies, carried out by Orosz (2009) in a Hungarian EFL context, found that primary school children were able to learn as many as 1,000 *lemmas* in a year, and knew about 3,500 out of the 5,000 most frequent *lemmas* after four years of primary school study. Another study by Hirsh and Nation (1992) looked at novels for teenagers or young readers – the same audience aimed at in this investigation – to establish the ratio of unknown to known words to allow successful guessing of unknown words. The results showed "that under favorable conditions, a vocabulary size of 2,000 to 3,000 words [would provide] a very good basis for language use" (Hirsh; Nation, 1992, p. 9).

While the knowledge of 3,000-word *families* is typically reckoned a bare minimum by Cobb (2007), Meara's studies (1995) suggest that a basic vocabulary of 2,000 words would account for about 80 percent of what people saw or heard. It can then be argued that L2 learners who reach that plateau would have a realistic level of language competence, being more capable of understanding the important meanings carried by words and able to see patterns in the way those words behaved. This seems to be in tandem with other studies which claim that a 2,000 vocabulary would enable L2 learners to start working with authentic language and authentic materials in another language. However, vocabulary growth for EFL learners is unlikely to reach the annual growth target of 1,000 word families "that can reasonably be expected of most

young L1 learners, but learning 500 word families per year [could] be an attainable target as long as the EFL learners receive the monitoring and support they need in order to achieve this goal" (Webb; Nation, 2017, p. 66).

Research in the depth and breadth of English acquisition in the recent past has shown the advantages of a contextualized exposition to language in the classroom. One approach to vocabulary learning that has been widely recommended by researchers is the selection and use of different materials on the same topic with words grouped semantically into lexical sets to increase the potential for the recurrence of exposure of target vocabulary (Nation, 2020). Investigations show that the relationships between the words that are learned together have an impact on learning. This leads to the assumption that teaching collocations can be very beneficial. It can be a very effective approach but not in the initial years of Elementary school, the context of my investigation. In Chapter 3 – Methodology – the groups of words selected by the software, also called 3- and 4-grams, will be mentioned and dealt with, even though they are not necessarily considered collocates. If a target vocabulary set consists of words that often appear together in sequence, this might have a positive effect on learning (Tinkham, 1997). Teachers would agree that the more learners read, the more high-frequency vocabulary they could encounter and possibly learn, and the easier the reading and understanding would become.

In sum, key words grouped semantically and groups of words which appear more frequently together, generated in context by software, can then be the basis of the pedagogic corpus for this investigation. The resulting number of words will complement the number of words in general English that is being taught simultaneously to the young learners. Number which is estimated by different scholars to be learnt yearly.

2.2 Word lists

Despite acknowledging the convincing sound arguments in favor of a contextualized exposition to L2, as well as the fact that much of L1 vocabulary growth is the result of repeated encounters with words in context (Webb; Nation, 2017), Meara (2001, p. 2) claims that *word lists* still have an important role to play in the acquisition of a new language. He argues that this role is particularly important at the beginning stages of learning an additional language. He clarifies his point by stating that

> the reason for this is quite simple. When you first start to learn a new language, the biggest problem you face is that you can't recognize any of the words.

> Nothing that you see or hear in the new language makes any sense at this stage, because all the words are unfamiliar.

Meara (2001, p. reinforces his argument in favor of word lists by claiming that

> young children learn their first language by acquiring single words in the first instance. They eventually get round to putting these words together into phrases[15] and sentences, but this development takes a long time. Children don't start using two-word utterances until they have a basic vocabulary of about 100 words.

The first core vocabulary list, West's General Service List (GSL),[16] was compiled as early as 1953 and has since proved its usefulness for both teaching and testing purposes. "The rationale that underpinned the GSL was that a lexical repertoire consisting of the most basic 2,000 words of English would be a good basis for learning English as a foreign language" (O'Keeffe *et al.*, 2007, p. 46). A plethora of *word lists* have been compiled since the 1900s, under an array of different criteria for their organization to cater for the teaching methods' demands of the time. They have mainly tried to include the most relevant and useful general English lexis for vocabulary learning, high-frequency samples that are supported by research findings. The idea behind developing the lists was that they should represent the higher frequency end of a learner's vocabulary. Therefore, the development of specialized vocabulary lists (Coxhead, 2000; Ward, 1999) would suit the goals of this study and be used to meet the learning goals of learners of English as an additional language[17] as well.

Due to the dynamic shifts in L2 teaching methodologies to contemplate better practices in the teaching-learning environment in recent decades, and with the increase in the computing capacity to aid the researchers, more complete *word lists* have also emerged. With the advent of specialized digital tools to replace much of the manual annotation work of the past, nowadays corpora contain billions of words. *Word lists* provide a shortcut to the improvement of performance in all skill areas, as learning a large proportion of the words that we are likely to encounter increases the potential for comprehension (Webb; Nation, 2017). They maintain that decontextualized vocabulary learning exercises are useful and often

---

[15] Chunk and lexical bundles (n-grams) come in handy at that point in the learners L2 development along the *Ensino Fundamental I* grades. However, the study did not focus on collocates as such, it just used groups of 3 and 4 words that were generated together by the software in some of the activities.
[16] The list was compiled by Michael West and substantially revised in 1995 by John Bauman and Brent Culligan.
[17]Available at: https://iaoed.org/downloads/prac06e.pdf.

> very effective at enabling learners to link form to meaning. Exercises such as those involving the use of flashcards and the keyword technique tend to result in fast and efficient gains in knowledge of the form – meaning connection of words, so they might provide a useful starting point for the development of lexical knowledge that could then be expanded through encounters with the words during listening and reading activities. They provide a quick way of developing a lexical foundation that can be used to make learning with meaning-focused input easier (Webb; Nation, 2017, p. 238).

The authors also state that "learning the higher-frequency sets of words within the lists may provide the greatest benefit to learners" (Webb; Nation, 2017, p. 196). As Laufer and Nation (2012, p. 171) put forward, "at the beginning stages of language learning, […] a wordlist could be a lexical syllabus […] as the words on the list could serve as a target for word-focused practice." Lists are undoubtedly a great source of authentic language as long as they are used well by teachers in preparation of pedagogical tasks. Provided teachers use such corpora-informed syllabus appropriately, it could give learners a solid foundation (Meunier; Reppen, 2015). Additionally, Reppen (2010) indicates that frequency lists and work with KWICs in corpus-based inquiries can be very useful for vocabulary instruction at the entry point of language learning.

One of those lists, highly recommended for beginner learners of English, is the most recent one The Essential Word List[18] (Dang; Webb, 2016 a) which includes sets of words (624 content words grouped into 12 sub-lists of 50 words, one sub-list of 24 words followed by a list of 176 function words) that are ordered according to their frequency. Two other well-known lists and to-date widely used in elementary schools in English-speaking countries are the *sight word* lists, one of them compiled by Dolch (1936) with 220 words and the other by Fry (1980) with 1,000 high-frequency words. Those words are used so often in print that together they make up an estimated 75% of all words used in books. Some of them cannot be decoded using conventional strategies[19] so memorizing them until they are known by sight is beneficial.

Sight words are known by sight, whose identification is triggered in memory very rapidly (Ehri, 1992). The process at the heart of sight word learning is a *connection-forming* process, i.e., connection between the sound and form of words since they do not follow the traditional word formation process.[20] In almost every school in the US, K-3[21] teachers assign their students these words to study and learn because they are the most frequently-occurring

---

[18]See: https://elt.oup.com/teachers/hvil/?cc=br&selLanguage=pt.
[19] There are many, depending on the country where English is taught. There are different techniques, such as combining sounds and groups of letters, using the phonics method, for example.
[20] It refers again to the association of sounds to letters and / or groups of letters.
[21] K-3 focuses on students in kindergarten through 3rd grade.

words in children's texts. "Sight-word knowledge provides a scaffold of understanding and confidence for new readers who need to read with understanding. In language pedagogy, any word that is read sufficiently often becomes a sight word, one that is read from memory (Farrell *et al*., 2013, p. 1). When sight words are known well enough, readers can recognize their pronunciations and meanings automatically without any attention or effort at sounding out letters (Laberge; Samuels, 1974).

The last argument above made us realize that the approach to teaching `sight word` words could very well suit the teaching of key words from the word lists of the pedagogic corpora. Those words would not have anything in common regarding their pronunciation and form. As the words would be chosen for their frequency, the understanding and memorization process would occur after repeated exposition and use in context. All the reasoning mentioned corroborated our choice to compile pedagogic corpora, with words grouped semantically in two school subjects, generated in lists of frequency in concordance lines. Those corpus-informed lists are the core of this study.

2.3 Words in a corpus

Conversely, and despite their usefulness, those lists include only single words which can be argued that the learning and retention of decontextualized words might not be as effective as if they are introduced to learners in their contexts. This aspect was initially investigated by many researchers, among those Lewis (1993) and later by Tinkham (1997) when analyzing the learners' exposure to words that often appear together in sequence. A phraseological approach to exposing learners to "multi-word units, many of which are as frequent as or more frequent than single items which everyone would agree must be taught (O'Keeffe *et al*., 2007, p. 46)" can be more relevant in the learning process.

Attempting to address the aims of the study and contemplate content word variables, such as degree of frequency, regularity of its combinations with function words, this is a corpus-based investigation where corpus is understood as "any body of text that is any collection of recorded instances of spoken or written language" (McEnery; Wilson, 2001, p. 197). This methodology enables the researchers to build specialized corpora, as well as classify the content words and groups of recurrent word combinations, also called lexical bundles (Biber *et al.*, 1999) or n-grams (Banerjee; Pedersen, 2003), according to its frequency and appropriacy. According to Biber *et al*. (1999), frequency is a fundamental characteristic that defines the word clusters. While by no means the only criterion, the basic idea is that "frequency of form and

meaning is the most reliable predictor of what can be most usefully taught at different points in the learning process" (Cobb, 2007, p. 479), including the early stages.

The orthodox view on corpus size is that the larger, the better (Sinclair, 1991). However, a corpus may be more specialized and quite small, for example, containing only 50,000 words of text, or very large, containing many millions of words (McCarthy, 2004, p. 1). Thus, it can be affirmed that the specialized corpora which are purpose-built, level-appropriate to be presented in this study will be considered to represent a smaller slice of language. Once available, hopefully in a user-friendly dedicated platform, teachers of English and of other subjects will be able to resort to them to collect information and plan interdisciplinary activities in English to motivate learners to be more involved in some areas of the *Cross-curricular Contemporary Themes*[22] (*Temas Contemporâneos Transversais* in BNCC).

One purpose for the use of corpora in language teaching is that it brings authenticity into the classroom. Another one is that it enables materials designers and teachers to examine what language learners are exposed to and develop more effective and better-informed pedagogical materials and tasks. Not only do corpora make it possible to expose learners to authentic language, but they can actually present them with a large number of authentic instances of a particular linguistic item. It seems that language noticing (Schmidt, 1990), frequency of occurrence (Ellis, 2012) and recurrence of exposure (in concordance lines) have pivotal roles in the L2 learning literature. Manipulating concordance lines can help learners see qualitative patterns of use beyond frequency. When particular structures that were given prior exposure are used again, syntactic priming occurs, making way for the learners to reuse them productively according to their needs. Language intake can then be identified and the learning process continues. This *condensed exposure* (Gabrielatos, 2005, p. 10) can, among others, contribute to vocabulary expansion and retention, and heightened awareness of language patterns (Granger, 1998). The combination of explicit instruction (Ellis, 2002) with target language (TL) recurrence of exposure (Gabrielatos, 2005) can promote noticing (Schmidt, 1990) of words and multi-word sequences (Cortes, 2004).

After determining the many variables which directed the investigation paths, it is relevant to state that the word lists extracted from the corpora were the basis for the activities we created for the treatment lessons. This is going to be fully explained in the methodology Chapter 3.

---

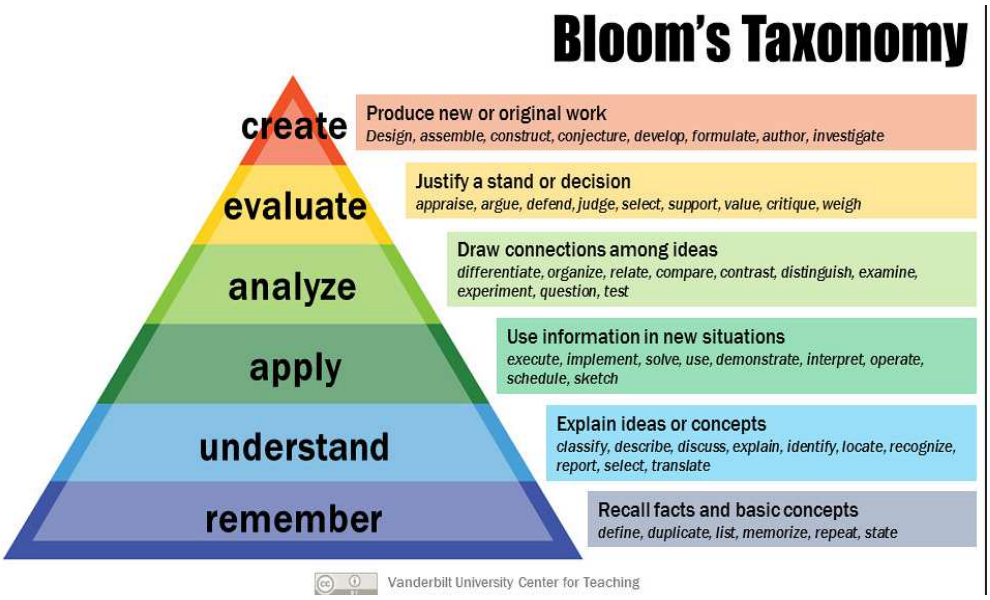
[22]See: http://basenacionalcomum.mec.gov.br/images/implementacao/guia_pratico_temas_contemporaneos.pdf.

2.4 Catering for learners' cognitive and linguistic skills

In 1956, Bloom *et al.* published a framework for categorizing educational goals: *Taxonomy of Educational Objectives* (*apud* Armstrong, 2010). This framework, widely known as Bloom's Taxonomy, has been applied by generations of K-12 teachers and college instructors in their teaching. It consists of six major categories: Knowledge, Comprehension, Application, Analysis, Synthesis, Evaluation and their subcategories. In their seminal work, the categories after Knowledge are presented as *skills and abilities,* with the understanding that knowledge is the necessary precondition for putting these skills and abilities into practice.

Language is the use of sounds, grammar, and vocabulary according to a system of rules that is used to communicate knowledge and information, whereas the word *cognition* refers to the process or act of obtaining knowledge through not only perceiving but through recognizing and judging. Cognition also includes such thinking processes as reasoning, decision-making, categorizing, selecting, problem-solving, recalling (remembering) among many others.

Table 1 – The six tiers of Bloom's Taxonomy



Source: Vanderbilt University Center for Teaching.

After the analysis of the six tiers, or categories, of Bloom's taxonomy above, the approach chosen to be used in the lesson delivery in this investigation was the Data-driven learning (DDL). It activates many of those processes while requiring the knowledge of additional skills from teachers and learners to deal with the software and corpora (Johns, 1991). Processes like identifying, noticing, recognizing, categorizing, relating, contextualizing,

creating were considered in the theoretical rationale for the design of tests tasks and classroom activities to be described in Chapter 3.

Redecker (2017) suggested that DDL has been used to open up learning to new, real-world contexts, involving learners themselves in hands-on activities, scientific investigation or complex problem solving. The ideal user of DDL is motivated to investigate abstract meanings from patterns of language use in the corpus, and ultimately, store these patterns so that they can form part of their repertoire of language, which can be expanded over time (O'Keeffe, 2021a). Thus, the lessons were delivered in a combination of approaches such as the paper-based, hands-off (Boulton, 2012, p.1) soft version of DDL and the Content and Language Integrated Learning (CLIL) approach. DDL encourages the use of authentic materials, promotes learner-centeredness through real exploratory tasks and activities to raise young learners' awareness of target language aspects and patterns by noticing them consciously (Schmidt, 1990). While CLIL refers to situations where subjects, or parts of subjects, are taught through a foreign language with dual-focused aims, namely the learning of content, and the simultaneous learning of a foreign language (Coyle *et al*, 2010; Solves, 2018). Principles underlying this approach "refer to the fact that CLIL is believed to help achieve individual as well as educational, social, and intercultural goals for language learning" (Richards; Rodgers, 2014, p. 119).

Due to the non-availability of software for the young learners in the six classrooms (Chapter 3), it was agreed that they would receive the material, mostly printouts of concordance lines, from the teacher who accessed the computer and printed everything, hence the label *paper-based* and *hands-off*. This soft version differs from the hard version when all participants have general ICT competence and are autonomous in their searches, using the computer tools to identify and analyze patterns in the language – inductive discovery.

To cater for the linguistic skills, the teacher provided a supportive environment to the young learners in the classrooms to foster social interactions and exchanges of children's contributions when they worked in pairs or groups. This kind of supportive structure, or scaffolding, a term coined by Vygotsky (1962) who believed that language developed primarily from social interaction. His observations of children and adults engaged in different types of interactions led him to postulate that scaffolding could help learners make the most of the knowledge they already had and even move to a higher language level known as zone of proximal development (ZPD) as a positive result. His views had great influence in the understanding of how second language learning developed (Lightbown; Spada, 2013) in past decades. Scaffolded learning plays an important part in CLIL. Gibbons (2002, p. 10) defines it

as “the temporary assistance by which a teacher helps a learner know to do something, so that the learner will be able to complete a similar task alone”.

These assumptions tie in very well with the DDL approach and its focus on learning language in concordance lines generated by a software. It was used in the delivery of lessons during the treatment with the tasks carried out by the young learners in all classes.

2.5 Data-driven learning (DDL) and Second Language Acquisition (SLA)

After describing the benefits of introducing the language with a DDL approach, it was used with the activities in the classroom tasks and tests. For lower-level learners, DDL can play a role in scaffolding development of noun phrase patterning and usage as learners are encouraged to discover patterns of language, and in doing so, more complex cognitive processes such as making inferences are fostered. O’Keeffe and Mark (2022, p. 1) argue that “DDL can bring an acceleration of language frequency experiences to the learner, through a type of *input flooding* (after Sharwood Smith, 1993)”. It is a process of inductive learning which “implies a level of active participation [of learners] in the learning process [and] learner and teacher interaction with the corpus itself” (Chambers, 2010, p. 345). In addition, Gabrielatos (2005) and Leel (2011) specify some underlying concepts the DDL approach can harness together: constructivist theories of language learning, the communicative approach to language teaching, developments within the area of learner autonomy (Chambers; Kelly, 2002) and *grammatical consciousness-raising* of the language (Rutherford, 1987).

According to O’Keeffe (2021), “many have called for connections to be made between DDL and SLA (Flowerdew, 2015; Johansson, 2009) especially via a *usage-based* (UB) *model of acquisition* which is seen to align well with the DDL approach (Ellis, 2012; Römer, 2019; Pérez-Paredes, 2020; O’Keeffe, 2021). Many scholars look at SLA through the lens of the UB models that view language as being acquired through the exposure to and the use of language (O’Keeffe, 2021a; Meunier, 2020; Pérez-Paredes, 2020). The core tenet of the UB perspective on language acquisition is that our knowledge of language comes from experiencing and using it as part of a communicatively-rich human social environment (Ellis; Larsen-Freeman, 2006). The UB posits that the mind acquires *constructions*, routinised patterns of form and meaning and holds that the more often they encounter a particular construction, or combination of constructions, the more entrenched it becomes (Langacker, 1987). UB evidence suggests that the process of learning an additional language involves intentional pattern finding which develops along a cline from basic formula (word combinations) to slot and frame sequences to fully abstracted constructions (Ellis, 2003; Pérez-Paredes, 2020; Mark; O’Keeffe, 2022). This

exposure and use promote constructivist usage-based learning (Crosthwaite, 2022), placing noticeable emphasis on the child's ability to create networks of linguistic associations (Lightbown; Spada, 2013). Likewise, DDL has a focus on guiding learners towards regularities so as "to make them aware of generalizations in patterns of form and meaning" (O'Keeffe, 2021).

The great majority of corpus linguists in Brazil have developed studies that focus on higher education and, in some cases, related to English for specific purposes (ESP). But not one contemplated the same aims or age group of the participants our investigation had. Corpus-based studies have been developed at PUC-SP (e.g. Berber Sardinha, 2017, 2021) and UNESP in São Paulo and in the countryside (e.g. Pinto *et al.,* 2021; Pinto; Garcia, 2022), and many Federal Universities also in Minas Gerais (e.g. Almeida *et al.*, 2023; Dutra *et al*., 2022; Costa, 2020) and Rio Grande do Sul (e.g. Bocorny *et al.*, 2021; Bocorny; Welp, 2021), for example, have renowned graduate linguistics programs that do corpus-based research.

Books with tailor-made designed activities based on corpora mainly to students at the higher education level have been launched more recently. A few exceptions in the Brazilian context with corpus-based design materials have aimed at high school students (Pinto *et al.*, 2023) and junior high school (Tartoni, 2012). She focused on the use of DDL in a 9th grade classroom, where learners worked with 'to' and 'for' as keywords in concordance lines, and Oliva (2018) who had her students work hands-on with editing tools and DDL in the academia are some examples of most current investigations. Pinto *et al.* (2023) organized a series of DDL-focused lessons on items of the language system: grammar, discourse, pronunciation, and vocabulary, for elementary to advanced learners, to be delivered hands-on and hands-off.

Despite the above publications, DDL with younger learners (9 – 12 years old) is still an under-researched area of Applied Corpus Linguistics worldwide. In a recent publication, Boulton (2020, e-book loc., 346) stated that although "[he has] been researching data-driven learning for many years now, [he knows] virtually nothing about DDL with younger learners. None in a primary school context". Most implementations of the DDL approach have been at the intermediate or advanced level (Boulton, 2008), a few with secondary learners (Pérez-Paredes, 2020; Crosthwaite; Stell, 2020; Wicher, 2020) while the number of DDL studies on primary-age data "can probably be counted on one hand" (Crosthwaite; Stell, 2020, p. 150).

2.6 DDL and pedagogic corpora

The above-mentioned fact that the use of DDL is under-researched among primary school young learners may be due to the fact that not only "teachers themselves may lack the

necessary digital literacy to use existing corpus tools, [but also they] may not always see the added value of integrating DDL in the prescribed curriculum" (Meunier, 2020, e-book loc., 880). Teachers' roles change fundamentally when working with DDL,

> as [they are] no longer the sole source of knowledge about the target language, but rather a facilitator of the learning process, helping the learners to interpret the data, and giving them advice on how best to search the corpus and analyze their search results (Chambers, 2019, p. 354),

"while the investigative work is carried out by learners and has been compared to the work of Sherlock Holmes" (Johns, 1997, p. 101). The more autonomous work will be responsible for the different degrees of accomplishment and empowerment learners will be able to experience. The teachers are still invaluable as curators of the information readily available, but their roles have been shifting to being more supportive to learners as the new generation gradually takes agency of their own learning path. It is a long-held belief among educators, that contemporary education should enable learners to be more engaged and committed to their own learning process and responsible for the results (Chambers, 2010). Thus, this is the window of opportunity teachers have to motivate learners to make effective use of the digital tools available to improve learning and lighten their weight as linguistic authorities they traditionally have had (Aston, 2007).

As mentioned before, barriers to the use of DDL include "very little time for extra professional development of activities in-service pre-tertiary teachers appear to have" (Bingimlas, 2009); also the lack of software for specific use with young learners (YLs) and a lack of corpus literacy reported for pre-/in-service teacher trainees of YLs. In order to implement DDL in a young learner beginner-level classroom, to my knowledge not attempted yet to this date, a challenge will need to be addressed first and foremost, namely the building of corpora of target language at an appropriate level for the students (Pérez-Paredes, 2020). Traditional corpora of authentic native speaker language are simply far too difficult (Anthony, 2013). To embrace it, two main resources are needed: the corpus aforementioned and a concordancing software to exploit it (Gilquin; Granger, 2020).

From a DDL perspective, Aston (1997, p. 13) states "[the] work with small specialized corpora can be not only a valuable activity in its own right, as a means of discovering the characteristics of a particular area of language use, but also an instrument to help and train learners to use larger ones appropriately." Aston (1997, p. 5) lists some advantages of working

with specialized pedagogic[23] corpora: " they are easier to manage, more fully analyzable, easier to become familiar with, easier to interpret, to construct, to reconstruct, are more clearly patterned and their limits are clearer". Notwithstanding, the proponents of the use of DDL argue that it should not be the sole component in the language teaching work in the classroom, but rather an enhancement with corpora-informed work. It will add variety to a lesson and raise learners' motivation towards learning.

2.7 Digital literacy for teachers and learners

This investigation aimed at addressing the demands of contemporary society for a more inclusive classroom, claiming that both teachers and learners should be better equipped in the digital environments. However, even though the software to be described in the next section was used only by the researchers due to the lack of available hardware, we believe it is relevant to discuss here the reasons why it was chosen and the advantages teachers would have if it were available.

As an immediate result of the fast pace of change in the educational scenario, the sudden advance of technology has brought into the scene the need to know English to navigate and visit websites, as it is the most commonly used language on the internet. As of 2018 and onwards, the guidelines in the *Base Nacional Comum Curricular* (BNCC)[24] state that teachers need to be skilled and equipped to help young learners acquire and develop, among other competencies, the digital competency in the *Fundamental* I grades.

The wakeup call started more than 10 years ago when Godwin-Jones (2015), and a few others before him like Reppen (2010), claimed that language instruction had begun to move ever more into online spaces, requiring more knowledge and skills from teachers to facilitate their students' learning. There were claims that more focus was needed to be given to the development of teachers' use of technology – ICT competence – that could support their own pedagogical work and core teaching performance (McKenney; Visscher, 2019). Recently, Crosthwaite (2022) corroborated those statements by positing that still today many teachers of young learners lack both the technical and pedagogical knowledge to integrate Computer Assisted Language Learning (CALL)[25] applications into teaching practice.

---

[23] Pedagogic corpora refers to a body of texts to be used in the classroom to support teaching (texts from the learners' coursebooks) with any additional texts that the teacher may bring into the classroom. It was coined by Willis (1998).
[24] Brazilian National Syllabus Core.
[25] The term Computer Assisted Language Learning (CALL) was coined by Hardisty and Windeatt (1989).

Therefore, it is of paramount importance to consider, first, the extent to which the teachers have access to the hardware and, second, if they have the ICT skills to be able to navigate the online medium. They also need to learn to curate the appropriate information before they can implement innovations in the classroom. In tandem with those demands, it should be acknowledged the existing paucity of hardware in many public schools (Mendes; Finardi, 2020) and even in most private schools, which may be one of the most demotivating obstacles to the digitization process in Brazil. Those requisites do reflect the current situation, a result of decades of little investment in hardware and maintenance. A common criticism is still that the changes require considerable investment in terms of training for teachers and learners to understand the rationale underlying the effective use of the digital tools. Additional drawbacks are the constant justification for the postponement of implementation of practical actions related to the beneficial use of technologies and the misinformation and prejudice against the use of computers in schools (Ribeiro, 2020).

The scenario described above reflects just some of the difficulties teachers have to face and overcome to transform their classrooms into 21st century educational environments. First, perhaps beginning with Google searches (Boulton, 2012), they need to learn to select the appropriate texts or group of texts which address their audience of learners. Second, teachers need to learn to use the digital tools to search for language patterns, and only then they would be ready to help learners interpret the findings and make effective use of them in the classroom.

In a nutshell, to start reducing the digital gap, pre-tertiary teachers should be exposed to the online medium during in-service and be instructed on how to navigate the web and on how to use the digital tools available to the benefit of their teaching. Teachers should be encouraged to step beyond their own comfort zone to help their learners develop their ICT competencies as well. On the other extreme of the continuum, learners also need to be digitally literate to look for information on the web. According to BNCC, they need to be exposed to the digital medium, learn to navigate safely, and identify suitable and trustworthy sites which suit their learning goals. They also need to learn to analyze critically the information they receive or send. Learners in the initial grades also need to learn how to use tools like Word and Images when copying contents with respect to copyright. They need to be acquainted with the use of web search engines appropriately to understand how to find what they need and be able to understand and interpret output of its particular discursive functions in context (Hafner; Candlin, 2007).

To empower learners with digital skills, the elements of motivation and challenge speak louder among young adolescents. And that is the moment, the window of opportunity, when

teachers need to be ready to stimulate learners to work with the concordance lines, to notice language patterns that are repeated and made salient and start doing their investigative analysis (Johns, 1991).

2.8 The concordancer #LancsBox 6.0

To access corpora data, it is paramount that teachers have the ICT skills to work with any of the many software available on the web, the concordancers. They are corpus analysis tools that search texts based on a word or phrase provided by the user and yield concordance lines which show the word or phrase in contexts, ranking them according to their frequency in that corpus. Despite its relevance in a corpus-informed investigation, we could not rely on enough hardware in the school to cater for all the YLs involved. As a result, the description of the software I am about to make is to justify my choice among so many options on the web and also raise readers' awareness of the advantages both learners and teachers can have in the future.

The digital tools need to be accessible, so that teachers can use them and instruct students in their use to improve language learning through self-regulated discovery work (Winne, 2017). Even if the learners have limited access to the digital world, teachers should start introducing elements such as problem-solving situations, queries that may trigger their interest and curiosity. Hendry and Sheepy (2022, p. 439) put forward that "learners can use corpus analysis tools to support vocabulary acquisition (1) as a reference to identify important words to study, (2) as a reference to check for patterns in typical usage in authentic texts," and language improvement and development of autonomous work. The literature suggests that if the digital tools are (1) hard to use or (2) perceived to be hard to use, then widespread adoption of the tools is not likely (Hendry; Sheepy, 2022). These authors mention the importance of the multidimensional construct of usability to identify and select the most appropriate concordancer to use in the classroom. According to the International Organization for Standardization (IOS), usability or ease of use, "is the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specific context of use" (2018).

Nowadays, more and more software are freely available to download from the web. Hendry and Sheepy (2022) in a recent study comparing concordancing software, present a very thorough analysis which point out that #LancsBox[26] was found to be the easiest for some to use. The 6.0 version has a straightforward interface and accompanying tutorials and can be very

[26] #LancsBox: Lancaster University corpus toolbox. Available at: http://corpora.lancs.ac.uk/lancsbox/index.php. Accessed on: March 10th, 2022.

suitable as a first step to those teachers trying to get acquainted with the current technological tools. Those factors have prompted us to choose it for this investigation considering the aims of its creators. At the time, Brezina and Gablazova (2018) declared that they were interested in improving learner vocabulary instruction through corpus analysis, mainly keyword and collocation analysis.

By using a concordancer with a readily-understood interface, meeting the criterion of being user-friendly, we have addressed the *user-friendliness* aspect mentioned by Frankenberg-Garcia (2012), an initial hindrance to the development of one's digital literacy. In this study, the tools *KWIC, Words, N-grams, GraphColl* and *Text* were used as instruments to identify, classify, analyze, and yield the most frequent target language which was used in the classrooms. The functions of each tool, the processes and the selection of vocabulary of interest are going to be described in Chapter 3.

The corpus toolbox #LancsBox has already many corpora embedded in its system such as American English, British English, BNC, Brown, LOB, English Literature, etc. Additionally, one can upload one's specialized corpus and use it like the pedagogic corpora used in this investigation: corpus for elementary levels on Geography (COREL-GEO) and corpus for elementary levels on Science (COREL-SCI) to be described in Subsection 3.4.

Chapter 2 reviewed the available literature underlying current investigations in many studies of corpus-informed applied linguistics. The variables to be considered during the design of the work to be implemented were described and the most suitable options for each of them justified. In Chapter 3, Methodology, the practical aspects of the process will be demonstrated, possible future uses of the tools, their implications in the classroom and outcomes are discussed.

Chapter 3 – Methodology

3.1 Introduction

In this chapter we describe the methodology of the investigation. We will specify the participants, the teacher, and the context. We will also describe the process of selection and compilation of texts to build the corpora COREL-GEO and COREL-SCI and demonstrate the use of the software tools to retrieve and analyze the data, the target language to be used in the treatment. The activities used in the class and tests will also be explained. The description is complemented by subsections on the lesson plans with the guidelines, the protocol to deliver the lessons, data collection and the teacher's comments on the delivery of lessons and tasks.

3.2 The context, the teacher, and the participants

The public school chosen for our investigation was a reference Elementary and Middle school connected with a Federal University. It is renowned for its excellence in children and adolescent formal education and its links with the university which afford support to relevant research in many other areas as well. It was a suitable venue for the investigation although in March 2022 it was under tight constraints imposed by the City Hall due to the 2-year pandemic, still in effect at the start of implementation of the study.

As a result of the gradual return to face-to-face classes, only the regular students, teachers and the officially employed staff had been granted permission to the premises of the school. The school did not allow access of strangers to the main building or classrooms. Postponing the work did not seem the best option then. Therefore, a possible solution to overcome this major hindrance was to invite a regular teacher to do the field work. She would advise the researcher on the selection of the vocabulary, on the designing of the tasks and tests and she would implement them in her classrooms while keeping contact with myself throughout. On the positive side, I could right away see an important benefit for having just one interobserver being responsible for the class work in all groups: the activities would take place in the well-known natural environment with learners focused on what they were already used to doing, i.e., listening to their own teacher's voice, instructions, and commands without any external distractions in class. In addition to that, the implementation would be very similar in all groups and the format of lessons would not be too different from what the teacher was used to doing: deliver interactive communicative lessons. The innovation was the introduction of the

target linguistic features collected from oral and written texts produced naturally and the tasks were based on data-driven learning principles. The teacher's main aim was to draw students' attention to the target key language in context (KWICs) and still work interactively practicing the language skills, mostly reading, writing, and speaking focusing on vocabulary and grammar.

The teacher, Clara,[27] although young, is a seasoned teacher who has been teaching children, adolescents, late adolescents and even adults in preparatory courses for specific English exams for the past 20 years. Clara has a degree in English, an MA in Applied Linguistics and is qualified in Teaching English as a Foreign Language (TEFL). In addition, she had already dealt with Corpus Linguistics in her Master's dissertation. These factors contributed to the reduction of possible external variables interference as she was the sole responsible for the classes management: application of tests, delivery of instructions when working in the classrooms and assigning homework to the students. Currently she is the teacher and coordinator of all 4th – 6th grades of *Ensino Fundamental*: two groups in 4th grade (4A-4B), two groups in 5th grade (5A-5B), and two groups in 6th grade (6A-6B).

All groups had two back-to-back English classes of 60' per week, which made it easier for Clara to implement a series of intertwined class work activities with the target language in concordance lines (KWICs) between pre and posttests. The investigation involved 147 children 9 to 12 years old and, according to Clara, the students had been exposed to English in weekly lessons since the 1st grade (Table 2), but had a long 2-year period of interruption (2020-2021) during the pandemic. The differences among the groups were not only the ages of the students but also the years of exposure to English prior to the study (explained in the dialogue below). When the lockdown started, online classes were not totally feasible for some time and we can assume that some of the previous English input might have been forgotten. Due to this gap, one preliminary assumption was that older students would have better results in the assessed activities due to the previous pre-pandemic exposure to English, which eventually was proved to be incorrect (Chapter 4).

[27] Clara is a pseudonym she uses for privacy.

Table 2 – Description of participants and distribution of groups by grade

Fudamental I	MODE	2 back-to-back classes per week	Number of students	AGE	Years of English classes
Group 4A	Face-to-face	1h 20 min	29	9-10	3
Group 4B	Face-to-face	1h 20 min	25	9-10	3
Group 5A	Face-to-face	1h 20 min	21	10-11	4
Group 5B	Face-to-face	1h 20 min	25	10-11	4
Group 6A	Face-to-face	1h 20 min	23	11-12	5
Group 6B	Face-to-face	1h 20 min	24	11-12	5
			147		

Source: the teacher at CP.

The first step to conduct the study was to have the necessary approval of the investigation by the Ethics Committee (Appendices A and B) and then the necessary permission to implement it in the premises (Appendix C). Once granted, the teacher and myself started the interactions about the project online due to the remaining concerns over the pandemic (2020-2022). As the sole interobserver, she was coached constantly during the investigation by the researcher who established a protocol with guidelines (subsections 3.5.1, 3.5.2, 3.5.3) for the delivery of all lessons to ensure it would follow quasi-identical procedures to ensure maximum reliability and external validity to the data to be collected. The online chats in Portuguese aimed at giving the researcher the overall profile of all the students. Some of the questions asked are presented here:

> **Researcher:** *1) Desde que série os alunos do CP têm aulas de Inglês? Ou seja, os da 6a. série têm inglês desde 2017, os da 5a. série desde 2018 e da 4a. desde 2019?*
> **Clara:** *2017 foi o primeiro ano do inglês na 1ª. série escolar. Assim sendo, nossos estudantes da 6ª. série foram os primeiros da escola a terem inglês desde a 1ª. série. Todos os estudantes da sua pesquisa (4ª., 5ª. e 6ª.) estudam inglês desde a 1ª. série (mas tiveram 2 anos de aula online) na pandemia.*
> **Researcher:** *2) Quantas aulas por semana? Todos os níveis têm o mesmo número?*
> **Clara**: *Duas aulas geminadas por semana, ou seja, um encontro semanal. No 1º ciclo (1ª., 2ª. e 3ª. séries, é a única disciplina que tem apenas um encontro semanal. Arte, educação física, orientação de estudos e outras disciplinas, têm maior carga horária.*
> **Researcher** *3) Qual a duração das aulas por semana?*
> **Clara:** *Duração total de 1 hora e 20 minutos de aula. Da 1ª. à 9ª. série é a mesma carga horária. Na escola, inglês, espanhol e arte têm a menor carga horária.*

When the lockdown started, online classes were not totally feasible for some time and we can assume that some of the previous English input might have been forgotten. Due to this gap, one preliminary assumption was that older students would have better results in the

assessed activities due to the previous pre-pandemic exposure to English, which eventually was proved to be incorrect (Chapter 4). Nevertheless, Clara was well aware that variations could occur during the study, mainly due to the differences in ages and profiles of the groups. Lightbown and Spada (2013, p. 37) claim that "cognitive maturity, metalinguistic awareness, world knowledge and attitudinal differences among learners from different age groups" can influence levels of language acquisition.

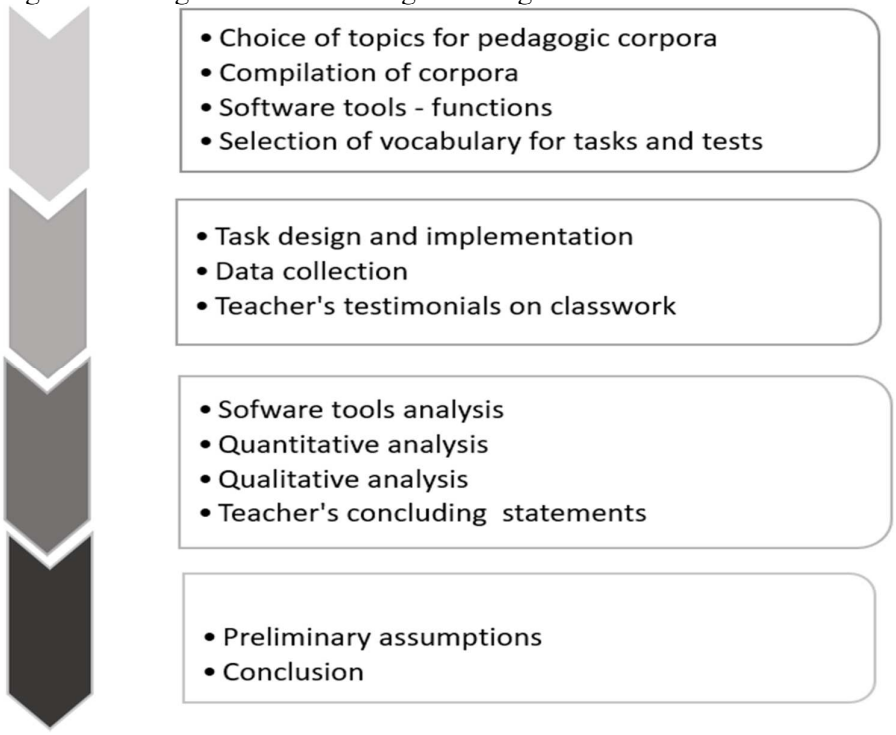
There were no structured questions at first, just open talks with Clara (the teacher) who asked for further clarification mainly related to classroom material, class management and data collection. Clara needed the assurance that she would be able to follow the guidelines in the Lesson Plans and would be able to deliver the instructions in English, as student-friendly as possible. That was also the moment we discussed the topics and the appropriateness of vocabulary for the groups, the type of tasks, the length of the tests and the inclusion in her regular lessons. As mentioned in the documents which granted the permission for the study, our main concern was not to interfere in the students' routine nor trigger any discomfort.

In the upcoming Sections, we will describe the research stages (3.3), the procedures to compile the pedagogic corpora and the concordancer tools used in the analysis of the subsets of corpora (3.4) as well as the implementation procedures and data collection (3.5). We will also explain the process of designing and implementing tasks and tests and include the teacher's impressions on the two 10-day periods of pre-tests, classwork and posttests.

3.3 Research stages

To carry out the investigation, a series of intertwined actions were taken before and after the data collection. They are specified below (Figure 1) and will be exemplified in more detail in Chapter 3 and findings and results in Chapter 4.

Figure 1 – Diagram of the investigation stages



Source: the researcher.

3.4 Compilation of pedagogic corpora: COREL-GEO and COREL-SCI

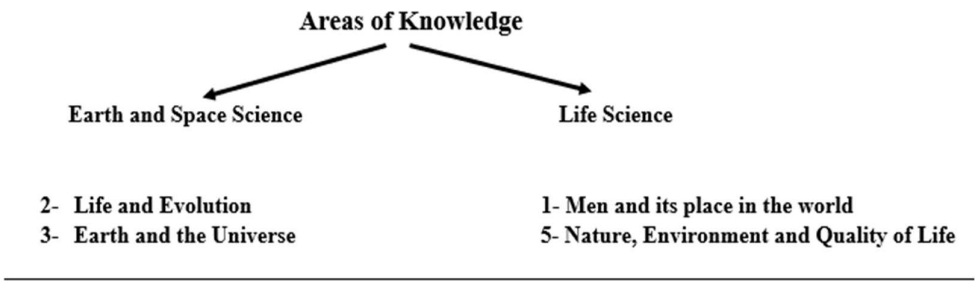
A corpus-based research process usually involves three main stages: corpus compilation, annotation, and analyses. The first stage is the most important when building the subsets of corpora. In order to investigate if vocabulary learning could be enhanced by corpus-based activities, first, we compiled the topic-informed corpora and uploaded them to the concordancer (subsection 3.4.1).

The first step was to build the corpora aiming at making them tailored to the recommended contents of 4th to 6th Geography and Science classes. This is in tandem with Pérez-Paredes's (2020) recommendation of a compilation of corpus data in English suitable for young learners. He defines pedagogic corpora (PC) as those which "follow design principles that differ from those present in corpora designed for research purposes: PC are topic-driven, they pursue pedagogic rather than linguistic representativeness, and they challenge traditional corpus-search behavior" (e-book loc., 2076-2077). While corpus size is usually an issue, it should be considered hand-in-hand with the appropriateness of its design. "In terms of suitability, however, it is often the design of a corpus as opposed to its size which is the determining factor" (O'Keeffe *et al*., 2007, p. 4). Therefore, I decided to compile *pedagogic corpora* as proposed by Willis (1998) as a body of texts to be used in the classroom to support

teaching with the addition of texts that the teacher could probably bring into the classroom (Gilquin; Granger, 2010).

To ensure the compilation would be feasible in the timeframe of the study, four major areas[28] were singled out for their cross-cultural foci and also for their overlapping of national with international primary school syllabi in English. Based on this initial criteria, two macro areas (Table 3)[29] in the domains of *Earth and Space* and *Life Science*, listed in the curriculum maps[30] of *Fundamental* 1 schools in Brazil were singled out.

Table 3 – BNCC macro areas of knowledge



Source: the researcher.

In the first selection, I chose the areas *Life and Evolution* (2) and *Earth and the Universe* (3). From the second selection, I opted for the areas *Man and its place in the world* (1) and *Nature, Environment and Quality of Life* (5). Subsequently, these areas were compared with international curricula through the analysis of a few American and Canadian elementary school curricula (available free online)[31] and two collections of workbooks[32] in English that indicate the corresponding school grades.

Crossing the information made it possible to identify an array of suitable themes or topics (Table 4) to build the corpora. We selected topics which were more universal, i.e., topics which would be found in most Elementary school grades regardless of their location. This is especially relevant as the vocabulary in the material had to be not only linguistically authentic but also graded according to the school 4th – 6th grades and meaningful to Brazilian learners. Thirty topics emerged as more recurrent in the sources and corresponding material was extracted from a small percentage of textbooks. The topics also informed the search and

[28] From the BNCC.
[29] The original source is in Portuguese and main areas have been translated to English to contextualize the choices made by the researcher.
[30] See: https://novaescola.org.br/conteudo/12720/bncc-baixe-em-pdf-o-e-book-de-competencias-gerais.
[31] Canadian: https://www.ontario.ca/page/ministry-education . American: https://ee.eanesisd.net/
[32] 180 Days of Science and Geography, Shell Education, K to 5th grade, 2014; and DK WORKBOOKS, Penguin Random House, Pre-K to 4th grade, 2016.

selection of material on the web to complement the desired diversity of trustworthy sources. In an attempt to motivate learners to engage in the classroom activities, the language contents would suit and benefit the different types of learners with different interests and different ages.

Table 4 – Themes / topics for the material

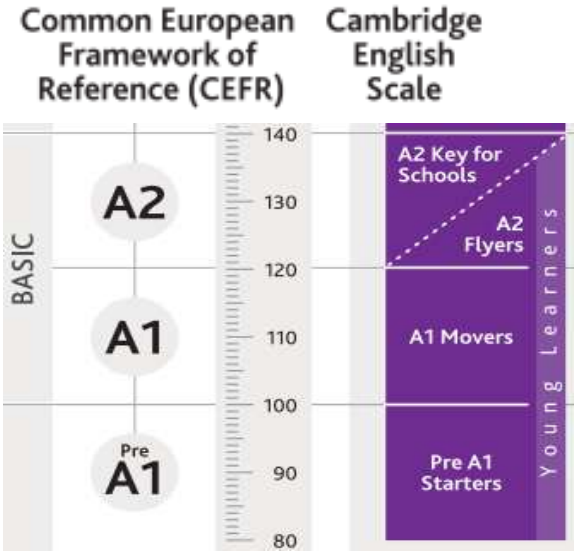
30 Topics	
Animals	Climate
Habitats	Natural disasters
Life Cycles	Seasons
Food Chains	Soil
Endangered Species	Erosion
Biomes	Climate change
Ecosystems	Resources
Plants	Renewable
Lifecycles	Non-renewable
Pollination	Fossil fuels
Germination	Pollution
Photosynthesis	Recycling - Conservation
Rainforests	Humans
Bodies of Water	Communities
Landforms	The Solar System - The Earth

Source: the researcher.

Regarding the web sources, it is worth mentioning that in 2020-2021, when we started the process of building the corpora described above, due to the worldwide pandemic and closing of schools, some web platforms[33] offered free access to their sources. This happened so parents and teachers could resort to their materials during the period when the vast majority of students were homebound. Most of the web written texts and video transcriptions on the topics were then checked against their linguistic complexity in the profiler vocabkitchen.[34] The texts had to be in authentic English[35] and in the appropriate linguistic level, namely A1 and some A2, entry points in the Common European Framework of Reference for Languages (CEFR), levels to be considered when working with beginner young learners (Table 5) in schools. The use of CEFR has been emerging as a standardizing measure for some time now, for example, in the design and compilation of new corpora (Tono; Díez-Bedmar, 2014, p. 165; Forsberg Lundell, 2021).

[33] One of them was: https://www.education.com/worksheets/ .
[34] See: https://www.vocabkitchen.com/profile.
[35] According to: https://www.pearson.com/languages, authentic material is any material written in English that was not created for intentional use in the English language classroom. [...] The best content to select depends on the learners, their level of English and the course content the teacher wishes to focus on. It's also a good idea to find out the learners' interests.

Table 5 – CEFR levels expected for young learners in schools



Source: English Language Assessment.[36]

Texts from different authentic sources were collected: printed workbooks A,[37] printed workbooks B,[38] web texts / video transcriptions and articles from assorted sources. For the corpora compilation, the workbook texts selected had to be scanned and saved in pdf.[39] The texts and the files with .docx[40] extensions had to undergo a transformation into files with a .txt[41] extension to become the final corpora. The process of transformation consisted in:

1) all files and texts were uploaded to Google Drive, opened and saved as Google docs in different folders;

2) the resulting *Google Docs* were then downloaded as .txt files and saved again in the new macro folders: COREL-GEO and COREL-SCI, separated in subfolders according to the 4th – 6th grades contents; and

3) the folders' contents, the 2 subsets of the corpora selected for the treatment in the classroom, were then uploaded to the concordancer #LancsBox 6.0 to be read, decoded and part-of-speech (POS) tagged.

The pedagogic corpora was built with a balanced number of 437 texts in Science (COREL-SCI) and 458 texts in Geography (COREL-GEO), 895 texts with 178, 669 words in total. Table 6 illustrates the sources and their participation in the overall number of texts:

---

[36] See: https://www.cambridgeenglish.org/exams-and-tests/cefr/.
[37] 180 Days of Science and Geography, Shell Education, K to 5th grade, 2014.
[38] DK WORKBOOKS, Penguin Random House, Pre-K to 4th grade, 2016.
[39] Portable Document Format by *Adobe Acrobat*.
[40] Microsoft extension document.
[41] Text Document file, a text document that contains plain text in the form of lines.

Table 6 – Information on COREL-SCO and COREL-GEO

Science COREL-SCI	Total number of texts	Total number of tokens	Total number of types	Total number of lemmas	Mean number of words
Written textbook	334	54,821	4,032	3,429	164,13
Written web	78	22,224	3,467	3,129	284,92
Spoken web	25	13,824	2,018	1,839	552,96
TOTAL	437	90,869	9,517	8,397	207,93
Geography COREL-GEO	Total number of texts	Total number of tokens	Total number of types	Total number of lemmas	Mean number of words
Written textbook	346	44,576	3,724	3,429	128,46
Written web	87	23,057	3,272	2,985	265,02
Spoken web	25	20,167	2,857	2,605	806,68
TOTAL	458	87,800	9,853	9,019	191,28
Overall	**895**	**178,669**	**19,370**	**17,416**	**199,40**

Source: #LancsBox 6.0.

Among all the 30 themes (Table 4), *Neighborhood* (Topic: Communities) and *Animals* were the ones singled out by Clara, the teacher,[42] as the most appropriate themes for activities to be implemented in all her 6 classes – 4th, 5th and 6th grades. The two topics would integrate seamlessly into their curricula. They were analyzed separately and stored in individual folders which will be easily accessible and made available in the future. The two subsets consist of 43 written and oral texts on *Neighborhood* and 63 written and oral texts on *Animals* in a combination of 106 texts in total (Table 7).

Table 7 – Topics (T1) and (T2) data

TOPICS	Total number of texts	Total number of tokens	Total number of types	Total number of lemmas	Mean number of words
Neighborhood (T1)	43	5,834	1,122	1,043	135,67
Animals (T2)	63	9,656	1,734	1,534	153,26

Source: the researcher.

Due to the diversity of themes, it is important to emphasize the need to incorporate the work with each one into a lesson after checking possible outcomes such as the students' interests and needs. Longer groups of new vocabulary in concordance lines may be demotivating for some or otherwise challenging positively depending on the activities in class.

[42] Clara's profile is described in Section 3.2.

3.4.1 The concordancer #LancsBox 6.0 - tools and their functions

Once the software is chosen, one needs to get acquainted with the short tutorials. Regarding the software #LancsBox, one needs to download it only once and leave it dormant on the desktop. When analyses of a corpus content are to be carried out, and if the corpus had been uploaded before, the software will perform the tasks demanded instantly. In this study, the researchers were the only ones accessing the software due to the constraints found at the *Fundamental* school. However, the word lists extracted from the corpora we made available to the teacher who participated in the whole process.

Figure 2 illustrates the software interface. It shows an array of user-friendly tools like *Words, KWIC, Ngrams, GraphColl and Text* in the header, the black bar at the top. These enabled the researcher to identify the most frequent lexis with *Words*, 3- and 4-word patterns with *N-grams*, number of occurrences of KWICs in the texts with *GraphColl*, and also the target language distribution in the files of the corpus. The tools also enabled the researcher and the teacher to select what was most appropriate for the tasks and tests (Chapter 4) carried out in the lessons.

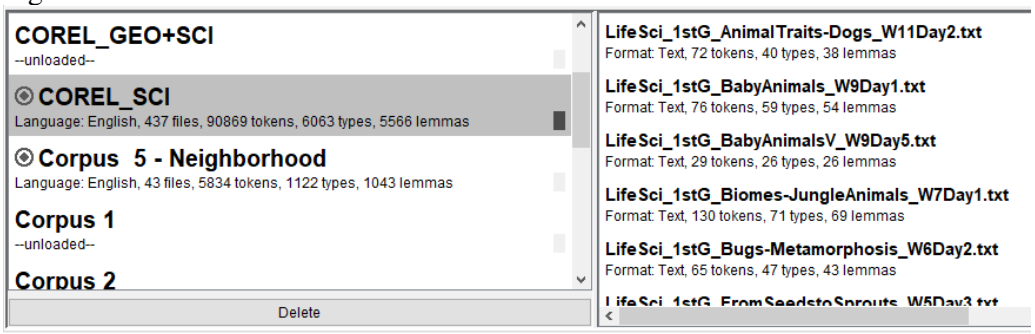
Figure 2 – #LancsBox 6.0 interface with the tools in the header

Source: #LancsBox 6.0.

In addition, *GraphColl* was used to show visually where the most frequent words could be found in the corpus, their frequency and distribution. Due to its visual appeal, the tool can also be more effectively used with learners in the classroom with computers available. Many possibilities for future hands-on use of the tools by learners will be mentioned in Chapter 4.

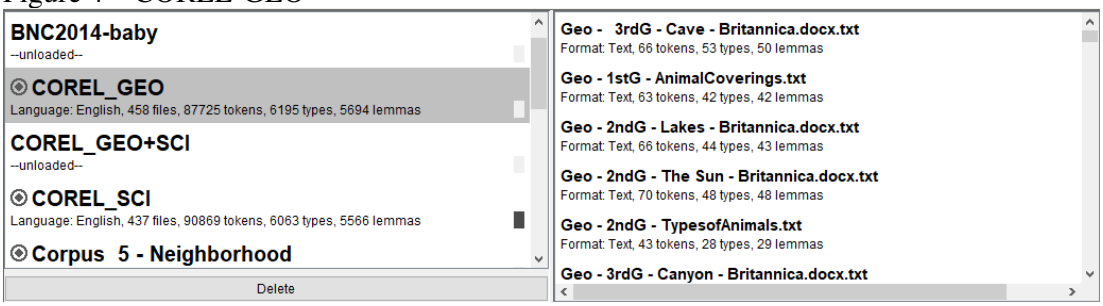
Below, the concordancer shows the two pedagogic corpora COREL-SCI and COREL-GEO uploaded already tagged (POS)[43] and with the main information highlighted in Figures 3 and 4:

Figure 3 – COREL-SCI



Source: #LancsBox 6.0.

Figure 4 – COREL-GEO

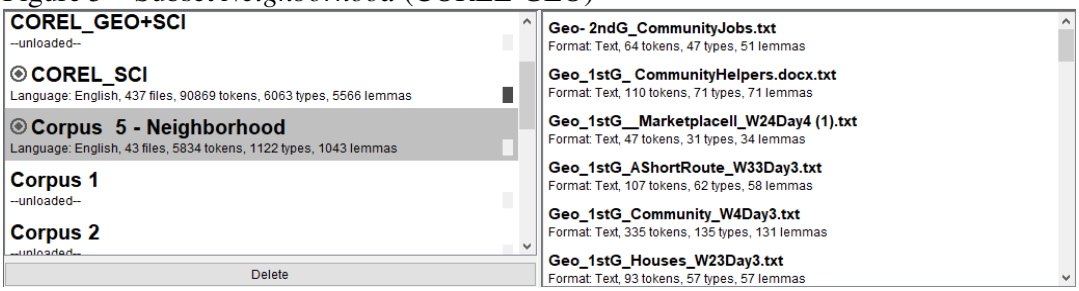


Source: #LancsBox 6.0.

However, as the investigation aimed at addressing only two specific themes, one subset of each corpora was chosen, *Neighborhood* from COREL_GEO and *Animals* from COREL-SCI were also uploaded and tagged. The corpora highlighted in Figures 3 and 4 above also display the subsets contents: the first with 43 files with 5,834 tokens and the second, with 63 files with 9,656 tokens. In the Figures 5 and 6 below, various corpora can be seen on the left side of the interfaces, but only the ones used in the investigation and highlighted in gray, have their files loaded on the right.

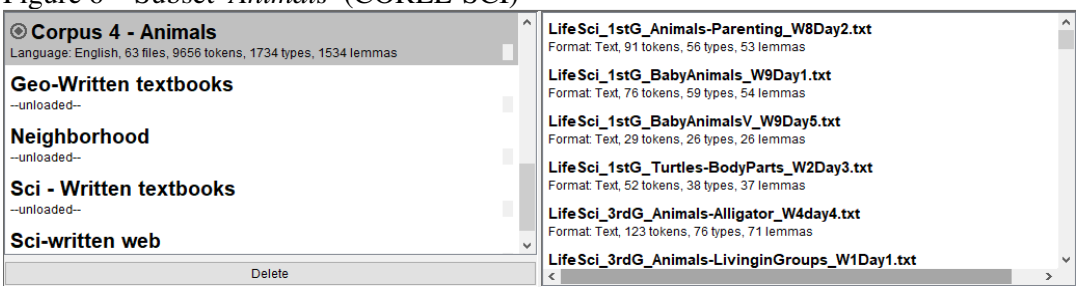
[43] Part of speech (POS).

Figure 5 – Subset *Neighborhood* (COREL-GEO)



Source: #LancsBox 6.0.

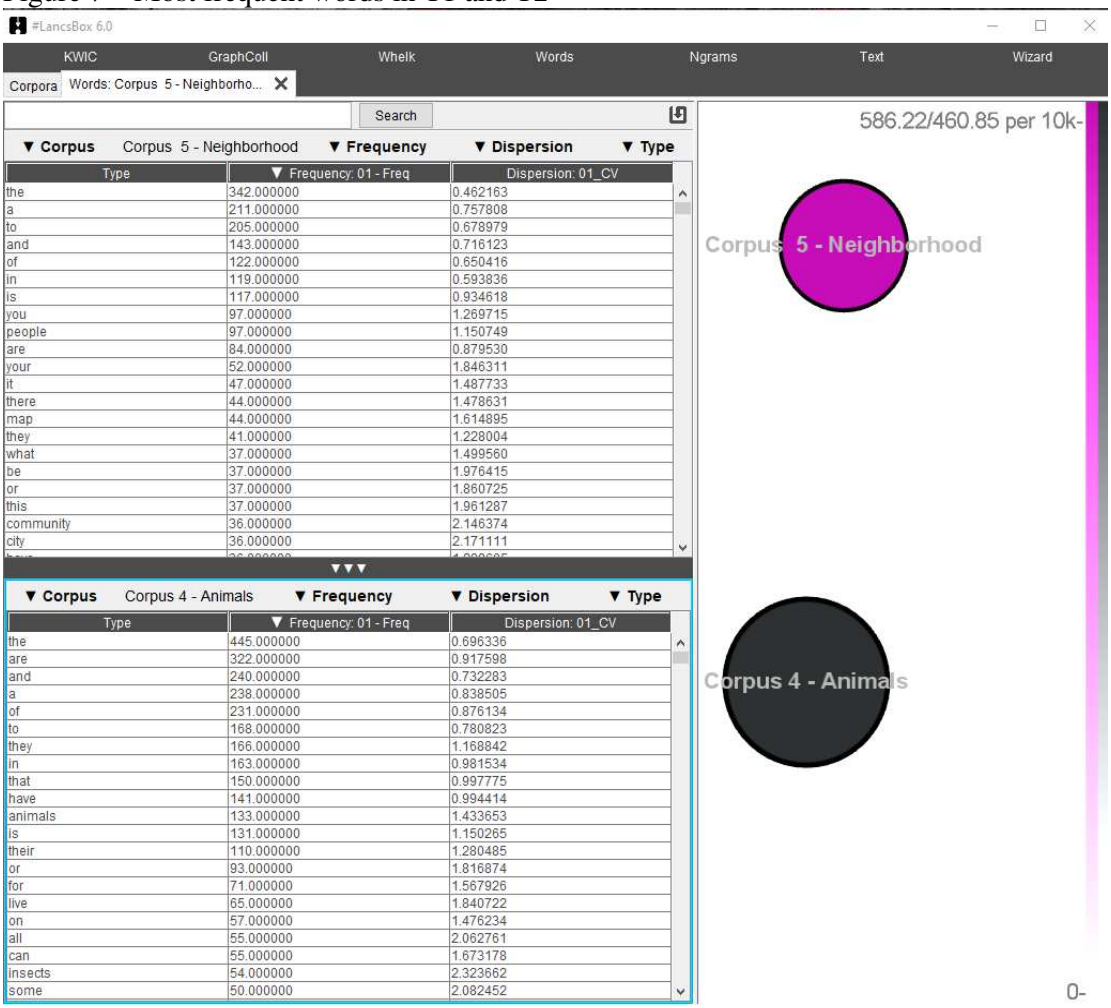
Figure 6 – Subset *Animals* (COREL-SCI)



Source: #LancsBox 6.0.

Once uploading both corpora, the first step was to check the topmost frequent words in each one and compare lists (Appendices L and M for longer lists). Biber *et al*. (1998) argued that the researcher should observe the data before deciding what and how they would work with it. According to his beliefs, most of the time the data lend itself to different types of analysis which should be noticed by the researcher so that the findings are meaningful and useful to learners. Observing the information retrieved in Figure 7, both tables reveal that the top words are functional words which are similar in both corpora. In our view, the outcomes point to the need to disregard them and focus on content words to identify the appropriate target language, content words in context.

Figure 7 – Most frequent words in T1 and T2



Source: #LancsBox.

Content words separated by classes will be listed to help the teacher and the researcher to select the most useful vocabulary for the learners. The findings will help the reader to understand why the use of technology provides state-of-the-art tools to enhance learning a language and how it can become an important and effective accessory to visualize patterns and clusters in context. Not to mention its role in aiding learner autonomy towards becoming the agent of his own language development.

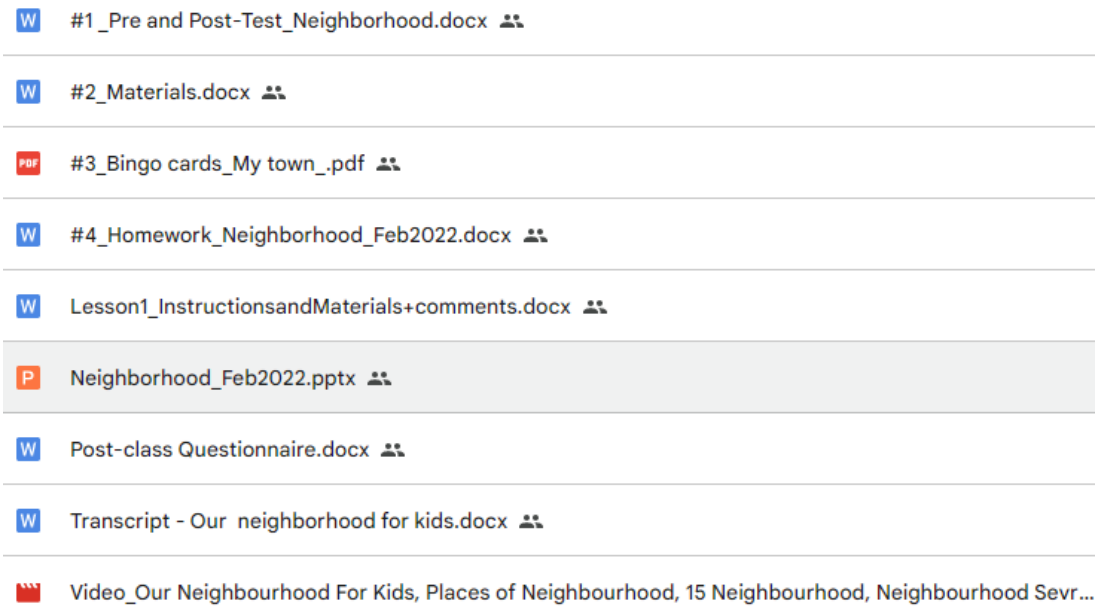
Chapter 4, Part A, will describe the utilization of the concordancer, outline the functions of its various tools, and present the obtained results.

3.5 Implementation procedures and learners' data collection instruments

As I was responsible for all the planning of lessons, the designing of classwork, homework, and the pre- and posttests, the teacher and I kept close contact throughout the whole

period. We engaged in a series of *Google Meet* chats and exchanged messages on *WhatsApp*. This process was undertaken not only to choose the most suitable topics for the learners but also to identify the target vocabulary (KWICs) along with its concordance lines and n-grams. All of these activities were aligned with the students' curricula.

Figure 8 – Screenshot of *Neighborhood* package contents – Topic 1 (T1)

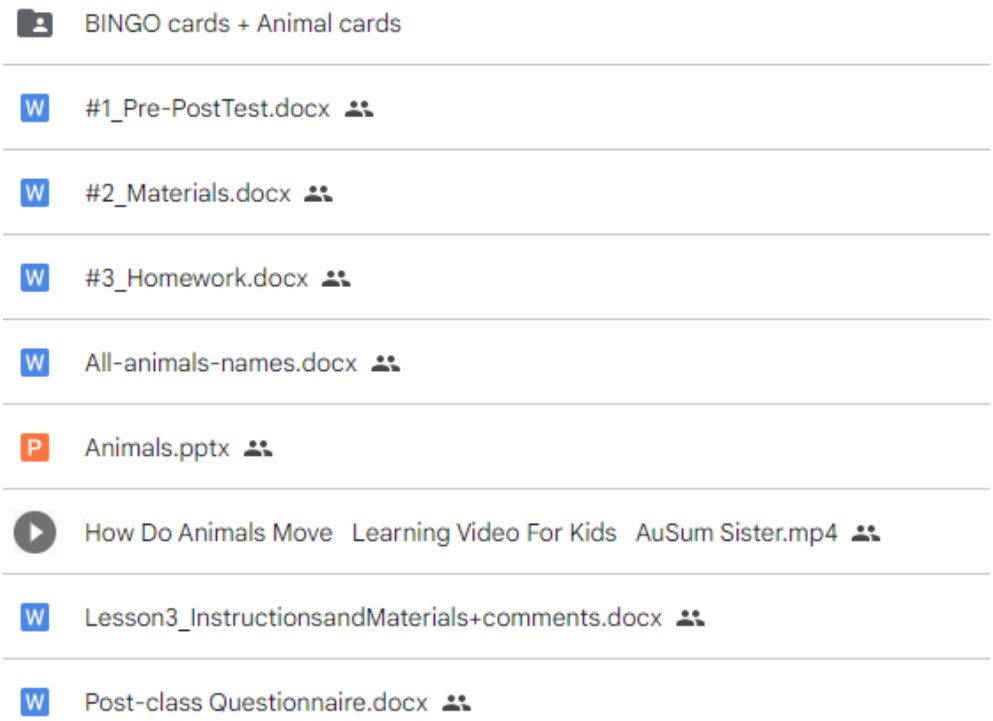


Source: *Google Drive* storage.

The continuous interactions aimed at assuring the target language had a direct connection with the interests and needs of the students so as to be meaningful, motivating and beneficial to them. Practicality and usefulness of contents for the students were aspects considered by both of us. The written guidelines as well as all the vocabulary in videos, texts or pptx were uploaded and saved in a shared Google folder (Figures 8-9).

The teacher received everything in a package for each individual topic containing all she needed in February 2022, a month in advance of the beginning of the school term in March. It would give her time to analyze all the activities and start planning the integration into the syllabus of each grade to start implementing it all in April. The design and methodological rationale underlying tests and activities delivered in the 4th, 5th and 6th grades were the same for Topic 1: *Neighborhood* and Topic 2: *Animals*. The processes will be further described in Sections 3.5.1 and 3.5.2.

Figure 9 – Screenshot of *Animals* package contents – Topic 2 (T2)



Source: Google Drive storage.

The implementation took place with the teacher using the DDL approach in a lesson delivered in a CLIL format in all the classrooms. They were delivered following the same sequence suggested in short lesson plans for each topic (Tables 9-10), and the pre- and posttests administration followed the same protocol, taking place within a 10-day interval. To ensure further impartiality on the results, all data collected in all classes were coded, corrected and assessed by myself following the same criteria. The tasks in the tests were quasi-identical to guarantee internal and external validity[44] of the results (Campbell; Stanley, 1966).

To design the pre- and posttests tasks as well as the classroom tasks, a list of target content words was prepared under the teacher's recommendation. These were analyzed and selected by the teacher to best suit the learners' needs and interests. Once again, we were reminded that traditional corpora of authentic native speaker language are simply far too difficult for beginner learners to grasp meaning of the words (Anthony, 2013). The work with the new vocabulary was only possible because the corpora had been built within the A1-A2 framework of language and also because they referred to topics that had all the materials and

[44] Validity refers to whether a test measures what it aims to measure. See: https://www.cambridgeenglish.org/blog/what-is-validity/.

tasks interconnected. I believe this is the most relevant aspect for having pedagogic corpora in an EAL class to improve learning effectively.

3.5.1 Designing the tests

The pre-tests (Appendices D-E) were designed to determine what vocabulary students already knew or could identify from previous contacts with the language. The results were compared later with the results of quasi-identical tests used as posttests. The tasks were sequenced according to the following learning strategies: language activation, brainstorming, contextualization, identification, and language production. The pre-tests and posttests carried 27 points each while class and homework together carried 16 points, which shows a distribution of 70 points. However, the production activity in each of the tests carried only 3 points in each one. The points were later normalized to 100 so that the statistics could reflect the treatment *de facto.*

Activating[45] prior knowledge (Table 8) of a language means both eliciting from students what they already know and building initial knowledge that they need in order to understand the meaning of the language. The activation (Task 1) will certainly foster appropriate brainstorming (Task 2), an activity used to generate ideas in small groups (Richards, 1990). In addition, contextualization (Task 3), "the fact or process of considering something in its context […] which can help in understanding it",[46] is tested, followed by identification (Task 4) which can be understood as recognition of language instances. Task 5 – Production, is the last task and it tests students' recall of the language they had been exposed to and had worked with in the class and home activities.

Table 8 – Underlying rationale for Tasks

Pre-test tasks: paper-based - 5 activities without the teacher's help

TASK 1	**TASK 2**	**TASK 3**	**TASK 4**	**TASK 5**	**Grades total**
Activation	Brainstorming	Contextualization	Identification	Production	
6	3	6	9	3	27

Post-test tasks: paper-based - the same type of activities without the teacher's help

TASK 1	**TASK 2**	**TASK 3**	**TASK 4**	**TASK 5**	**Grades total**
Activation	Awareness	Contextualization	Identification	Production	
6	3	6	9	3	27

Source: the researcher.

[45]The activation strategy was first coined by Harmer (2003) in his book "How to teach English" when proposing the ESA (Engage - Study - Activate) methodology.
[46]See: https://dictionary.cambridge.org/us/dictionary/english/contextualization.

The constructs underlying the design of each activity are labeled in Table 8 above and are identical in TASKS 1, 3, 4 and 5. In Task 1 (Figure 10), in both tests, students were exposed to the names of the buildings and had to match labels and buildings before attempting to do TASK 2. TASK 2[47] *brainstorming* in the pre-test (Figure 11) is the same that assesses *language awareness* in TASK 2 - posttest (Figure 12), because it was done after the class and homework. Below are examples of some of the tasks designed for both tests which illustrate the underlying constructs defined and the homework with the focus on the structure *there be – there is there are* – can be checked in Appendix F.

3.5.1.1 Pre- and posttests – Topic 1 (T1)

Below are samples of some of the tasks the young learners had to do in the pre- and posttests: Figures 10, 11 and 12.

Figure 10 – Pre- posttest - *Neighborhood* - Task 1 - *Language activation*

Source: the researcher.

[47] The labels followed a logical sequence of learning and teaching language communicatively but both tasks tested producing language.

Figure 11 - Pre-test - *Neighborhood* - Task 2 - *Brainstorming*

2- Choose **3 buildings and write sentences about them:**

a- _____

b- _____

c- _____

Source: the researcher.

Figure 12 – Posttest – *Neighborhood* – Task 2 – *Language Awareness*

2- Read the concordance lines below.
Then, choose **3 buildings (exercise 1) and write sentences about them:**

- There are many different types of map. This is because we use different maps for different reasons.
- find the best road to your friend's house
- find your way around a nature park
- City block with homes and stores
- Do you have friends in your neighborhood?
- Tall apartment buildings where many people live
- They may have museums, libraries, and parks.
- People live, work, learn, and have fun close to one another in cities
- They may be able to walk to school, the post office, the library, and stores.
- They may also use public transportation to get to different parts of the city.
- Tall apartment buildings where many people live
- There is a library nearby
- When you have a picnic at the park, you clean up after yourself
- A park map, for example, help you plan

a- _____

b- _____

c- _____

Source: The researcher.

Figure 12 illustrates Task 2 in the posttest with some examples of the concordance lines they had worked with in class. Clara explained that some students claimed they did not know how to spell some words they wanted to write and so she decided she would add some of the concordance lines in the test but without calling their attention to them. The analysis of the results in Chapter 4 will demonstrate it did not make a noticeable difference in the learners' written outcomes.

Figure 13 – Pre- and posttest – *Neighborhood* – Task 3 – *Contextualization*

3- Match the two columns:

(1) Hospital	( ) fireman
(2) School	( ) police officer
(3) Police station	( ) doctors and nurses
(4) Post office	( ) principal
(5) Bakery	( ) postman
(6) Fire station	( ) baker

Source: the researcher.

One important variable is the influence of classwork and homework in the posttest results. One of the tasks that shows it is Task 3 – *contextualization*. The vocabulary in the pre-test - Task 3 (Figure 13) was dealt with in the classroom when students watched a short video and worked with its transcription afterwards (Appendix H). Clara exposed them to the target language (KWICs) in context and worked with the text (Figure 14) to contextualize the neighborhood services.[48]

Figure 14 – Text related to *Neighborhood* services

107 Community Helpers People help out in the community. Bus drivers take people where they need to go.
108 Teachers teach students. Crossing guards help children cross the street. Fire fighters keep people safe when there is a fire.
109 Police officers protect people. Doctors help keep people healthy. Mail carriers bring mail. Farmers grow food for people to eat.
110 Vets help people keep their pets healthy. You can help, too. You can pick up trash. You can help a neighbor.
111 Who delivers mail to your house?
112 Who helps students get to school?
113 Who can help people in an emergency?

Source: #LancsBox - Tool: Text.

\

[48] On Neighborhood: NS LEARNING TOOLS, Our neighborhood for Kids, Places of Neighborhood, 15 Neighborhood, Neighborhood Services. Available at: https://www.youtube.com/watch?v=OQxRVOG10ZA.

Figure 15 – Pre- and and posttest – *Neighborhood* – Task 4 – *Identification*

4- Look at the picture about this town. Which buildings do you know? Write the names you can remember:

_____ _____ _____

_____ _____ _____

_____ _____ _____

Source: unknown on the web.

Task 4 above aimed at helping learners recognize information they were already acquainted with in the pre-test and could recall in the posttest. Moreover, Figure 16 below shows Task 5 – *Production.* In the pre-test, it aimed at raising the teachers's awareness of how much of the topic the learners already knew. However, Task 5 in the posttest raised expectations they could use many of the new words in personalized sentences to complete the activity.

Figure 16 – Pre- and posttest – *Neighborhood* – Task 5 – *Production*

5- Complete the sentences to describe your neighborhood:

. I live near the _____and _____

. There are _____

. There is _____

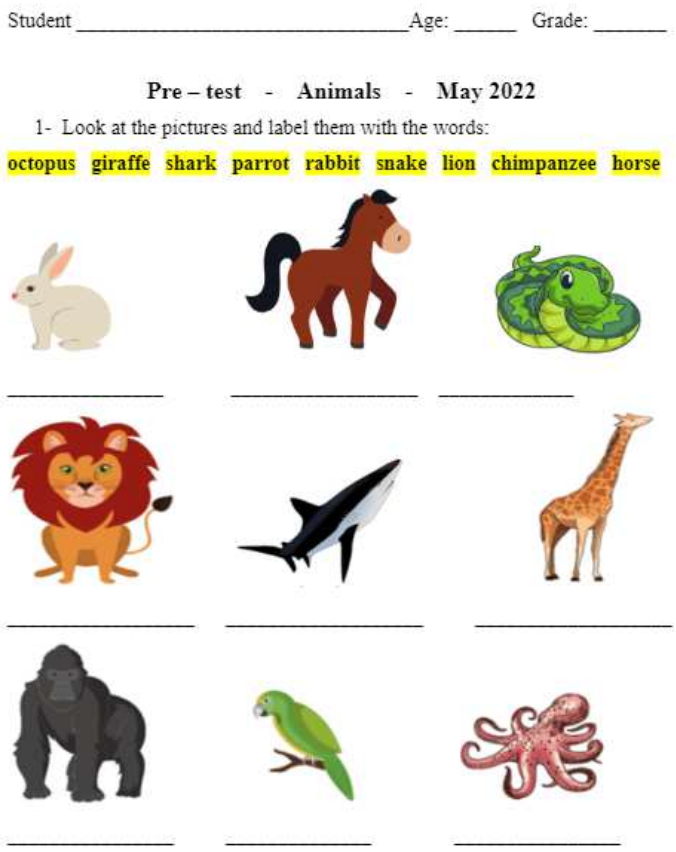
Source: see Appendix F.

After the classwork with the target vocabulary and the various activities with the concordance lines, the learners received the homework (Appendix F) handout not only to revise the names of buildings but also to consolidate the difference between *there is* and there *are.* It was also preparation for the posttest after an interval of 10 days.

3.5.1.2 Pre- and posttests – Topic 2 (T2)

The design and rationale underlying the pre- and posttests in Topic 2 – Animals followed the same principles used in Topic 1 (Appendices D, E). Constructs such as *identification, recognition, awareness, brainstorming, recall* and *contextualization* were contemplated and helped learners in the written work, the data which were collected and assessed.

Figure 17 – Pre-test – *Animals* – Task 1 – *Language activation*


Student _____Age: _____ Grade: _____

Pre – test - Animals - May 2022

1- Look at the pictures and label them with the words:

octopus giraffe shark parrot rabbit snake lion chimpanzee horse

Source: the researcher.

Figure 18 – Pre-test – *Animals* - Task 2 – *Brainstorming*

2- Choose **3 pictures and write sentences** **with the words you selected:**

Example: The **shark** lives in the ocean. The **shark** is white and black. **Sharks** eat fish.

a- _____

b- _____

c- _____

Source: the researcher.

The sentences produced by the students in Task 2, during *brainstorming*, were a demonstration of what they already knew about the topic. The same task in the posttest was aimed at showing how much of the input offered during the lesson and consolidated during homework would affect their *production.*

Figure 19 – Pre-test – *Animals* – Task 3 – *Language recall*

3- Complete **each group of sentences** **with the same word** from the list or with the ones you remember:

a) The _____ is a big cat.

The _____ lives in the wild in Africa and India.

The _____ hunts during the day.

The _____ have a beautiful mane around their head.

b) The _____ live in the water, in rivers and in a bowl in the houses.

Some _____ are gray, others are yellow, blue or red.

The _____ which live in the Amazon have a dangerous bite.

c) The _____ live on farms and on the plains.

Some children like to ride the domestic _____.

The _____ can be brown, black, gray and also beige.

Source: the researcher.

Figure 20 – Pre-test – *Animals* – Task 5 – *Language awareness and recognition*

5- Let's classify the animals. Read the names and put them in the right columns below:

- Tiger, dog, lion, fish, cat, cow, elephant, horse, snake, shark, frog, panda bear, giraffe, parrot, rabbit

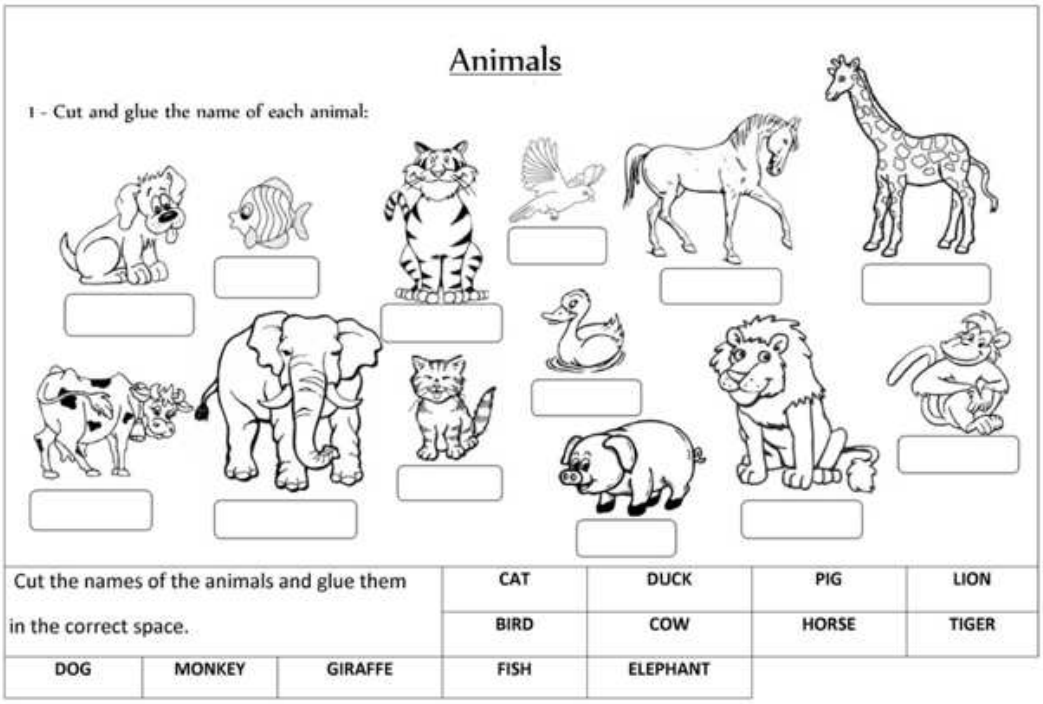
Domestic	Pets	Wild

Source: the researcher.

Tasks 2, 3 and 4 were aimed at guiding students towards producing their own sentences in Task 5 in the posttest. As for homework, the teacher asked for more demanding tasks in Topic 2 as she noticed the learners already knew most animals (Figure 21). It was decided that I would plan tasks which would require an expansion of vocabulary on the topic, but focused on the powerpoint work in the class and bingo game (Appendix I) with the animals and their characteristics. It is important to mention that they had worked with the concordance lines with that information as well. During the game they had to produce sentences orally about the animals they had on their cards to get the points. Examples: 1) The parrot is colorful; 2) The parrot has wings and flies; 3) Chicken cannot fly; 4) The tiger is the largest animal in the forest; 5) The lion has strong legs to run; 6) The monkey can jump; 7) Crickets are jumping insects; 8) Fish live and swim in the water; and 9) Sharks are dangerous.

Figure 21 – Homework – *Animals – Language Consolidation*

Label the animals and write sentences about **3 you like most: (0.5 each)**



Animals

1 - Cut and glue the name of each animal:

Cut the names of the animals and glue them in the correct space.			CAT	DUCK	PIG	LION
			BIRD	COW	HORSE	TIGER
DOG	MONKEY	GIRAFFE	FISH	ELEPHANT		

Source: the researcher.

1. Which animals do you know? (at least 3 - 0.5 each)

2. Which ones can you find in the zoo or at home ? (maximum 8 - 0.5 each)

In the Zoo	At home pets – domestic

3. Name and describe 5 of them. Say how they move (swim, fly, run or walk)

(0.5 each)

_____

_____

_____

_____

Source: the researcher.

3.5.2 Lesson planning

The 2-year pandemic brought a new understanding of the demands in the classroom environment, showing in practice that not only learners but also teachers need to be digitally savvy to stay abreast of the cutting-edge innovations and continue to be a player in the education field of work. It certainly begins with the introduction of innovations in the teaching approaches in the classroom as they can be the starting point to change the traditional standpoints learners and teachers have had in their relationship. In this study, the contribution for this change was the implementation of class material: authentic age-appropriate language from corpus-informed corpora in concordance lines intended to be interesting enough to keep learners engaged and working communicatively.

3.5.2.1 Brief rationale underlying the lessons' activities

Beginner language learners improve and develop their learning through many cognitive processes ignited by different techniques and materials used in language classrooms. One of them is copying chunks of language from models of texts they are exposed to in the lessons. Once the meaning is understood, the learner copies parts of the sentence, makes the changes they consider suitable, including the target language making the new sentence true to himself. In the past, *drilling* through substitution and completion work[49] were key features of the audio-lingual method (Skinner, 1957; Fries; Lado, 1979; Brooks, 1964). In the past, the technique contemplated only oral language but more recently it has been extended to written work as well. In this case, the emphasis should still be first on oral comprehension of meaning (Richards;

[49] A substitution drill is a classroom technique used to practise new language. It involves the teacher first modeling a word or a sentence and the learners repeating it. The teacher then substitutes one or more keywords, or changes the prompt, and the learners say the new structure. Available at: http://englishteachingtechniques.blogspot.com/2012/09/. Accessed on: August 20, 2022.

Rodgers, 1986) and only afterwards students would work with language inductively, trying to compare their production with the models given.

Nowadays, those techniques have had a comeback but under the contemporary notion of language being taught in context. This has been coined as *drilling in disguise* when the repetition *drill*[50] is carried out more communicatively, more meaningfully, and not mechanically as in the audio-lingual teaching method. The assorted variety of existing types of drilling is more creative and used in a learner-centered environment. It refers to the fact that "many communicative drills can be modified in the classroom to require meaningful communication" (Rubio *et al,* 2003, p. 18).

Having that in mind, for this investigation, DDL-type activities designed for the class work in the three groups of learners: 4th to 6th grades were delivered in a Content and Language Integrated Learning (CLIL) format since the language content was the focus and English the medium of instruction. CLIL refers to situations where subjects are taught through an additional language with dual-focused aims, namely the learning of content and the simultaneous learning of a foreign language (Marsh, 1994). In the pre- and posttests (Subsections 3.5.1.1 and 3.5.1.2) as well as the class work, the activities ensured cognitive processes[51] (Table 8, p. 60) were triggered by different techniques employed by the teacher to introduce and consolidate the target language. The activities which required substitution of words had the underlying rationale of affording identification and repetition of target language in different moments of the work.

The lessons were back-to-back and lasted 1 hour and 20 minutes each week. The DDL approach was not supposed to take the entire lesson, yet be a complement to the class routine, and used when learners worked with the discovery of new language with the concordance lines. The lessons followed a similar sequence with all the groups: language presentation (work with meaning at first through picture picture presentation or video watching), repetition of target language using the pictures, flash cards, or playing bingo, and a written activity. The innovation was the use of age-appropriate concordance lines[52] with the target language (KWICs) yielded by a pedagogic corpus which were copied and cut out as strips of paper to be manipulated by students, giving them the opportunity to work in pairs or groups to discover the lexis and language patterns inductively.

In the ideal environment originally suggested by Johns (1991), learners would analyze language yielded by software using the computer. This hands-on discovery work could not be

---

[50] Systematic repetition of facts or sentences to aid memorization (audio-lingual method).
[51] Different taxonomies of task types were used in task design (Bloom, 1956; Anderson; Krathwohl, 2001).
[52] Yielded by the software.

carried out during the investigation and so, the paper-based, hands-off soft approach was chosen by the researcher (Boulton, 2012). English was used most of the time and some translations occurred to clarify meaning. Clara gave learners all the required support during the lessons while following the suggestions in the Lesson plans (Tables 9-10).

3.5.2.2 Lesson plan – *Neighborhood* (T1)

Table 9 – Lesson stages[53] – Topic 1 – *Neighborhood*

Activities	Materials	Procedure suggested
Pre-test	Paper-based **First day**	Individual work without any help. Data collection
Classwork 1	*PowerPoint* slides	Awareness of key lexis attached to images they represent; work with 'word clouds'; personalization of buildings in one's own neighborhood; can be used at the end of class as well, in this case beginning with the video (4'15) + transcription.
Classwork 2	Paper-based Strips of paper	**Concordance lines work and n-grams** – strips of paper with the KWICS in context – in pairs. Models of authentic language use and usage. Work with 'substitution' of keywords to make it meaningful to each student.
Supplementary tasks	Video[54] + transcript or Bingo cards	To emphasize the KWICS orally and draw students' attention to them once more and prepare for the written homework. Flashcards can also be used.
Homework	Paper-based Tasks	Consolidation of KWICS; revision of structure: 'There is' and 'There are'. Students to access link and answer questions about the image: https://www.liveworksheets.com/gf787983nj.
Post-Test	Paper-based **10 days later**	The same as the pre-test; individual and without any help. Data collection.

Source: the researcher.

I devised all the activities for each of the topics with a DDL approach to encourage the teacher to add variety to her lessons, diversify her approach and also to encourage her to use corpora with her learners in a very learner-centered format. The centerpoint of DDL is the manipulation of the concordance lines for inductive work with the discovery of patterns of the word sequences and n-grams, to observe the target words in context (KWICs), identify and notice the surrounding words. The classwork material provided to the teacher was varied and Clara was free to use the slides of the pptx and short videos to introduce the topic, images to consolidate the words and associate them and bingo cards (game) for the recall and production at the end. The assorted material was produced or selected specifically for the groups which is

[53] The lesson plans and materials such as pptx or bingo and flashcards were designed and produced by the researcher.
[54] The video transcript used in the classroom is in Appendix G.

an idea corroborated by Frankenberg-Garcia (2012, p. 486), who argues in favor of "help[ing] language teachers to take their first steps in using corpora autonomously [to] encourage them to want to find out more about using corpora in the classroom".

The images in the *PowerPoint* slides used in Classwork 1 mentioned in the Lesson Plan above (Table 9) as well as the short videos were accompanied by word repetition aimed at linking the oral form to the images they represented, helping the process of memorization. Reppen (2010) corroborates the idea of working with sound and form at the initial stages of L2 learning and suggests a holistic approach using pictures and other visual aids to be matched to words. The flashcards were used mainly for recognition of form and pronunciation work. "Exercises such as those involving the use of flashcards and the keyword technique tend to result in fast and efficient gains in knowledge of the form – meaning connection of words (Meara, 2001, p. 238). As an example, the teacher would call students' attention to the context in which the word *neighborhood* occurred, i.e., to the words which came before and after *neighborhood* in the concordance lines and would replace the KWIC with another one of their preference. The aim was to draw students' attention to the word sequence, to raise their awareness to notice the language in use to enable them to recall it when writing their sentences.[55]

In another activity, students played with strips of paper (concordance lines) and created their own sentences or questions. Examples below:

1) concordance line*: There are many different types of maps. This is because we use different maps for different reasons.*
   Learner: *There are different types of medicines. This is because we use different medicines for different reasons;* and
2) concordance line*: Do you have friends in your neighborhood?*
   Learner: *Do you have friends in your school?*

[55] A more detailed plan of action in the classroom with T1 is in Appendix J.

3.5.2.3 Lesson plan – *Animals* (T2)

Table 10 – Lesson stages – Topic 2 – *Animals*

Activities	Materials	Procedure suggested
Pre-test	Paper-based **First day** **Docx* #1**	Individual work without any help. Remind them they will learn those words they still do not know in the following class. Names of animals, body parts and ways of moving should be dealt with in the DDL class. Data collection.
Classwork 1	*PowerPoint* slides	Awareness of key lexis – animals' names - attached to images they represent; the first 13 slides are very important; slides 13 and 14 are body parts; work with video + transcription. Key words: nouns and verbs but some adjectives may apply when they describe the animals. There are too many animals, but you can choose according to what you believe they already don't know.
Classwork 2 + Classwork 3	Paper-based Strips of paper **Docx* #2**	**Concordance lines work and n-grams** - strips of paper with the KWICS in context – in pairs. Models of authentic language use. As you know the learners well, they can work with two or more strips of paper, depending on how their understanding goes. Work in pairs or small groups to analyze the sentences and the KWICs neighbors. Work with 'substitution' of keywords and the surrounding words to make it meaningful to each student's preference.
Supplementary tasks	Video or Bingo cards	To emphasize the KWICS orally and draw students' attention to them once more and prepare for the written homework. Flashcards can also be used. This is the time for a fun game with a competition among different small groups of learners to make it easy for them to remember the animals' names. They can either guess names, or body parts always producing their own sentences. 'How do animals move. Learning Videos for Kids'.mp4 – AuSum Sisters[56]
Homework	Paper-based Tasks **Docx # 3***	Consolidation of KWICS; vocabulary expansion and additional variety according to their grades.
Post-Test	Paper-based **10 days later**	The same as the pre-test; individual and without any help. Data collection.

Source: the researcher.

The DDL approach in class and the techniques used were very similar to what happened with Topic 1. Same instructions for the activities, similar language presentation and similar media to give support to the sequence of activities. The data was collected and assessed by the researcher and then results submitted to tests to have quantitative information for comparison. The findings are described, and some assumptions made in Chapter 4.

[56] The video above has been discontinued from *YouTube*. However, I had downloaded it when preparing lessons and saved a copy of it.

3.5.3 Teacher's testimonials – remarks on the classroom work

During one of the first talks online, Clara mentioned that students asked many questions when handed in the pre-test in one of the groups, including requests for clarification of vocabulary. Clara said she even had to explain the instructions and say what was expected from them without giving them the answers. She also told them they could leave questions undone, unanswered, but some learners did not like it. She had to calm them down so they could complete the work. The comment below was made by the teacher after the initial phase of the implementation of the investigation and describes her students' reactions and her feelings towards the work with authentic material from the corpora:

> I collected all the sentences the students produced after the work with concordance lines. As this was the first group, I handed out different strips to different learners and many students did not understand parts of them. As a result, for the work in the following classes, I tried to choose 'concordance lines' which were more related to their previous knowledge and their current language context. I felt they were more motivated and produced more sentences.

After the initial concern with the work with concordance lines, and her changes in how she approached the students to do it more effectively, the lessons took place in a more relaxing way with students competing among themselves to produce sentences. She also mentioned it was quite interesting to see them work with the lines and select chunks from them to use in their writing. When she realized many of the students could infer the meaning of the KWICs and recall some of the target vocabulary to describe their own neighborhood in the written sentences, she finally felt it was worth it. In her words,

> the written tasks were done in class since I wanted to observe their difficulties, being able to help them when needed. I do believe everything learned in isolation is much more difficult for learners. Contextualized words and chunks make learning more meaningful and applicable for learners. Not all of them internalized the new n-grams well enough to use them, but the sentences were mostly meaningful mainly due to revising previous samples.

She mentioned she had never worked with specialized corpus-informed material before and felt that at times the input with concordance lines offered more challenges to the younger learners. Although unaware of what the investigation results would be at the time, the teacher was in fact instinctively trying to answer Anthony (2009, p. 1) who inquired, "there is no

question that the use of corpora in the classroom has value, but how useful is concordancing with beginner level EFL[57] students?"

After working with a KWIC in the classroom, for instance, *neighborhood* (Topic 1), Clara worked with some 3- grams to talk about her own neighborhood. Many of the samples illustrate the use of *find your way, need a map, different kinds of, in my neighborhood, a map can and live in a*. In another activity, Clara would use another approach to the KWICs to introduce the work and motivate students to build sentences with the 3-grams. She would elicit their reply to the question:

> *T: I live in a quiet neighborhood. I can see the mountains and a big valley around. And you?*
> *T: Now, tell me about your neighborhood. Where do you live?*

The teacher replicated the procedures used with Topic 1. She used a few texts with Topic 2 in the classroom to clarify the meaning of the KWICs or expand learners' vocabulary from other concordance lines. Clara used different approaches to call learners' attention to segments of the sentences that could be useful when they produced their own sentences to describe the animals in their homework and posttests. Some examples are *jump very well, jumping insects, animals that have strong legs to run, have feathers, … that cannot fly* and *cannot fly at all*. Learners worked with them while looking at images in the pptx , handling concordance lines like the ones below and playing bingo.

One last remark is that she noticed that the older the learners were, the less tidy their sheets of exercises or tests were. Those learners gave her the impression of lack of focus and perhaps, lack of interest. However, those who participated with attention produced sentences which were more creative and most of the time with new combinations of the vocabulary they had been working with (Chapter 4 – Qualitative Analysis).

3.5.3.1 Questionnaire for the teacher

After implementing the tests and the sequence of classroom tasks on both topics: Neighborhood and Animals in her groups, Clara was invited to answer a short structured written questionnaire (Table 11) below.

[57] English as a Foreign Language (EFL). It is used in the teaching - learning environment.

Table 11 – Post class-evaluation – Teacher's questionnaire

4th - 5th - 6th grades

The application in the classroom is underpinned by a tripod of constructs: Content and Language Integrated Learning (CLIL), information from a specialized pedagogic corpus and the approach Data-Driven learning (DDL):

		YES	NO
1	Did the specialized corpora-informed material make any difference in regards to learners' interest and engagement in the tasks when compared to your regular approach?		
2	When vocabulary is grouped in lexical sets is it easier for learners to relate to it and grasp meaning more easily?		
3	Was your experience using the **KWIC**s – key words in the concordance lines – with the learners positive?		
4	Were the learners motivated to do the follow-up task? Was it in class or for homework?		
5	Did they do it on their own or needed someone`s help?		

Source: the researcher.

The answers complemented the teacher's remarks on using the concordance lines and would guide the researcher to adjust the tasks to make learning more effective in future work with concordance lines. Since on different occasions Clara had had many opportunities to express herself mainly in relation to class management, the questions were prepared so that we could conclude her work with the groups. In tandem with the first delivery of lessons and administration of tests, Clara wrote short notes for herself about class management which later were shared. In addition to the classwork, she also photographed the students while working with the concordance lines to share with the researcher (Appendix K).

The qualitative and quantitative analyses of outcomes will be thoroughly described next in Chapter 4, which should yield evidence whether resorting to corpus-informed pedagogic corpora could heighten learners' expansion of English vocabulary in their early years of primary school.

Chapter 4 – Results and discussion of outcomes

4.1 Introduction

In this chapter we will present the results of the investigation distributed in two main areas: Part A – the relevance of compiling pedagogic corpora for younger learners and the language analysis of concordancer tools, and Part B - the results of learners' production in the posttests after working with the concordance lines in the classrooms. The two corpora COREL-GEO and COREL-SCI yielded topic-informed and grade-appropriate vocabulary which, once selected and dealt with by learners, can shed light on the effectiveness of implementing a new approach to language learning in the English classroom.

First, Part A, subsection 4.1, where we will show the concordancer tools functions and analysis results and the selection of the linguistic features to prepare activities. Discussing these results, we will be able to answer research questions (i) and (ii) which were presented in Chapter 1 and are reproduced here:

i) Which are the most frequent topic-related L2 content words and 3 and 4 n-grams (lexical bundles / chunks) in COREL-GEO for 4th – 6th grades? (Appendix L); and

ii) What are the most frequent topic-related L2 word combinations 3 and 4 n-grams (lexical bundles / chunks) in COREL-SCI for 4th – 6th grades? (Appendix M);

Second, Part B, subsections 4.2 and 4.3, where quantitative and qualitative analysis of the results of learners' language production will be presented in an attempt to answer research questions (iii) and (iv):

(iii) Can activities implemented with a Data-driven learning (DDL) approach expand learners' topicalized vocabulary to boost their progress in English?; and

(iv) Are the results significantly different from one grade to the others when the same tasks are worked with in the classrooms?

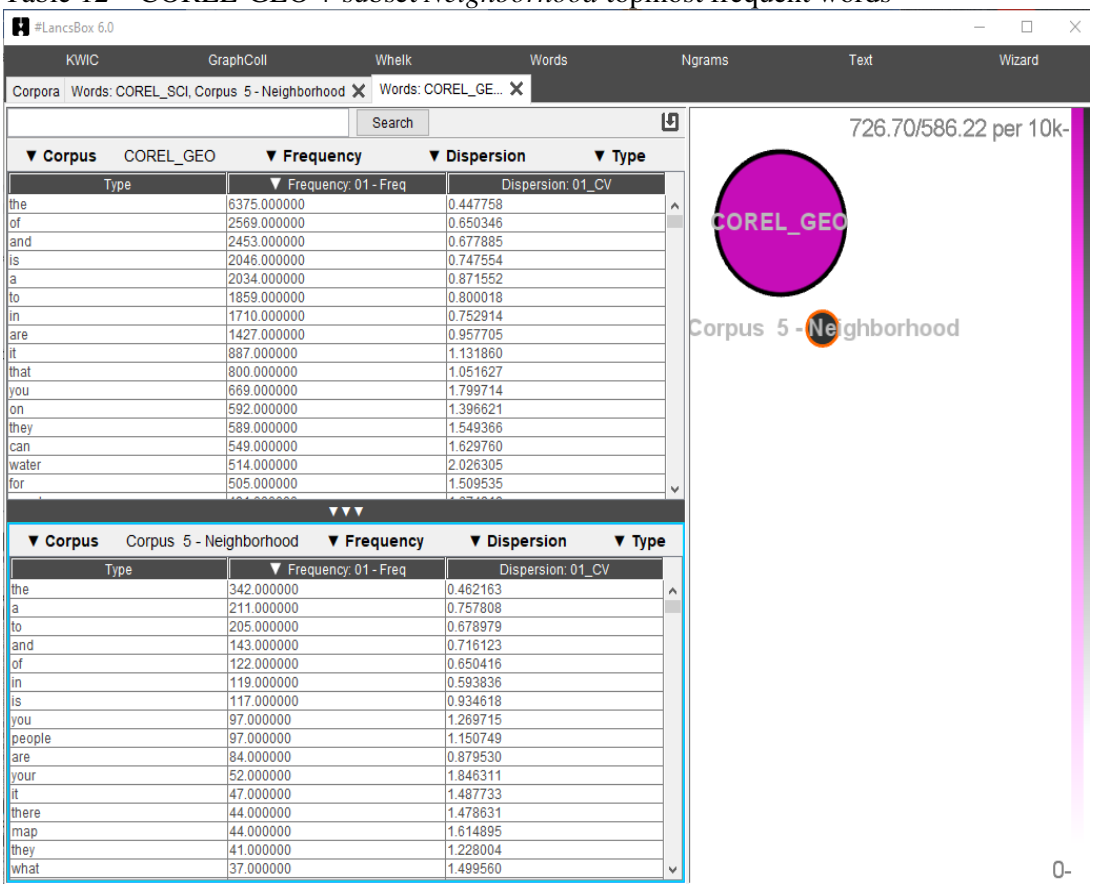
Due to the length of the *content words lists* in the pedagogic corpora, the complement to answers to questions (i) and (ii) above can be found in Appendices L and M. The justification for not including the lists in the body of the thesis is that the activities in the classroom and the production of the learners were reliant on only two subsets of those corpora: *Neighborhood* (T1) and *Animals* (T2). Information on those subsets can be found in Table 7 (Subsection 3.4) and Figure 7 (Subsection 3.4.1). The COREL-SCI and COREL-GEO corpora can be found in its entirety in Appendices L and M.

Part A

4.2 Concordancer language analysis – tools, functions, and results

This section explains how the concordancer tools carried out the different analyses of COREL-GEO and COREL-SCI contents described in subsection 3.4.1. After building the corpora, the second stage, annotation, and the third, analysis for later retrieval, were performed by #LancsBox 6.0, the web-based concordancer chosen for this investigation. Table 12 shows the most frequent words in COREL-GEO and compare with the topmost frequent words in subset *Neighborhood* (T1) while Table 13 compares the most frequent words in COREL-SCI with the top ones in the subset *Animals* (T2).

Table 12 – COREL-GEO + subset *Neighborhood* topmost frequent words

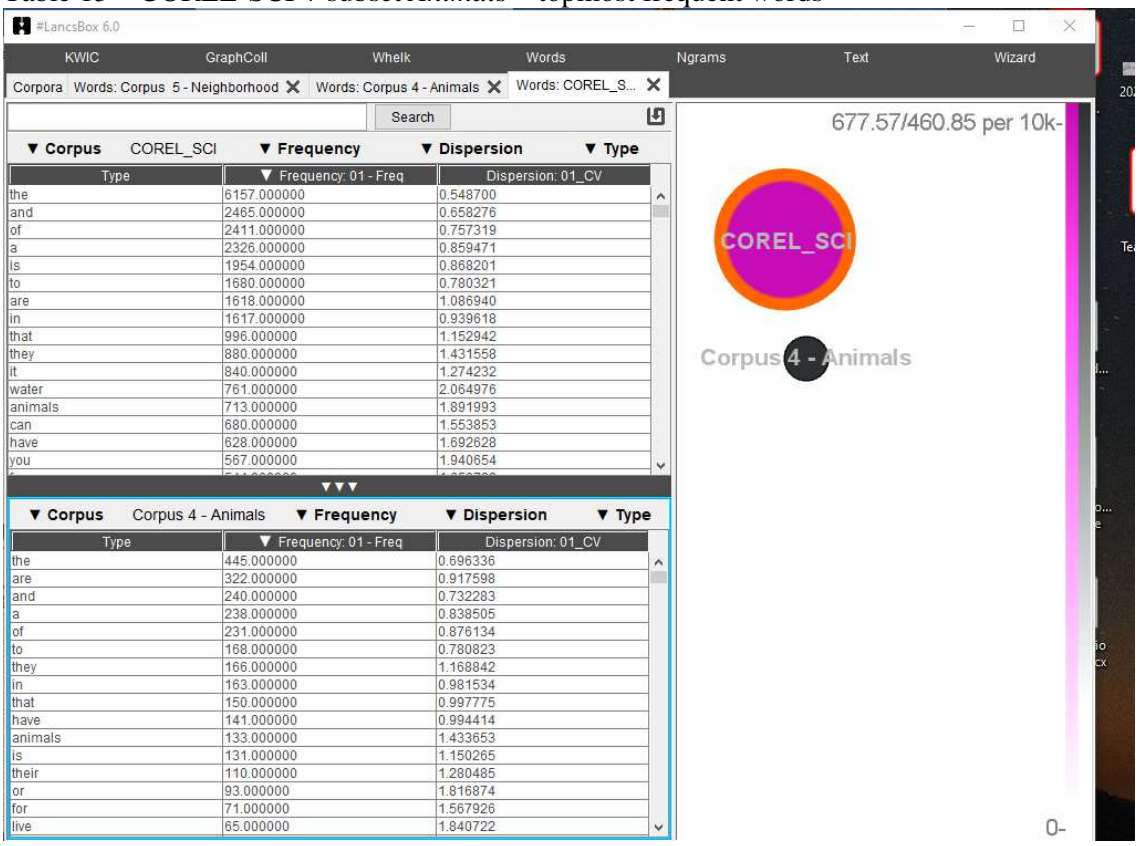


Type	Frequency: 01 - Freq	Dispersion: 01_CV
the	6375.000000	0.447758
of	2569.000000	0.650346
and	2453.000000	0.677885
is	2046.000000	0.747554
a	2034.000000	0.871552
to	1859.000000	0.800018
in	1710.000000	0.752914
are	1427.000000	0.957705
it	887.000000	1.131860
that	800.000000	1.051627
you	669.000000	1.799714
on	592.000000	1.396621
they	589.000000	1.549366
can	549.000000	1.629760
water	514.000000	2.026305
for	505.000000	1.509535

Type	Frequency: 01 - Freq	Dispersion: 01_CV
the	342.000000	0.462163
a	211.000000	0.757808
to	205.000000	0.678979
and	143.000000	0.716123
of	122.000000	0.650416
in	119.000000	0.593836
is	117.000000	0.934618
you	97.000000	1.269715
people	97.000000	1.150749
are	84.000000	0.879530
your	52.000000	1.846311
it	47.000000	1.487733
there	44.000000	1.478631
map	44.000000	1.614895
they	41.000000	1.228004
what	37.000000	1.499560

Source: #LancsBox.

Table 13 – COREL-SCI + subset *Animals* – topmost frequent words



Source: #LancsBox.

In a single glance it is possible to check the findings and find similarities in the results. Both subsets of both corpora reveal more function words at the top of the lists and very few content words. Therefore, to make the investigation more focused and meaningful to learners, it was necessary to have the concordancer generate and analyze the word classes: nouns, verbs, adjectives, and adverbs separately in subsets *Neighborhood* and *Animals*. More comprehensive lists of those subsets are in Appendices N – O.

4.2.1 Selection of vocabulary for the activities and tasks

In this section we will mention the concordancer findings to address questions (i) and (ii) to be found in Appendices L and M, narrowing down the information with the results of the most frequent word classes and n-grams in the subsets *Neighborhood* and *Animals* in Appendices N and O. We will also demonstrate how the content words were separated in word classes and n-grams, and then exemplify how some of the KWICs in concordance lines were selected by the teacher.

Once the subsets *Neighborhood* and *Animals* were compiled, the concordancer generated their most frequent content words. Surprisingly, some of the words among the 50 ranked as most frequent were not the ones both the teacher and I had expected to work with. The result was the same with the n-grams but the combinations we selected, unlike the individual words, were situated at a higher rank in the list generated by #LancsBox. As a consequence of that, other vocabulary scattered on the list was chosen to complement the material for the activities. An example of the lower ranked words were the names of buildings in a neighborhood. Although *hospital, school* and market were ranked in a higher position, *library, bakery, police station, fire station* were not. We assumed the reason for that is the fact the material in the subsets of corpora was aimed at native-speaker learners, and the vocabulary such as the *nouns* mentioned would not need to be in the youngsters' textbooks with such frequency.

Nonetheless, this fact did not pose any constraints to the task and test design as the *KWIC* tool could generate concordance lines with any word chosen, despite its degree of frequency. Our criterion was to select the KWICs which we thought would be meaningful based on the learners' curriculum maps. The KWICs which had more samples, and consequently would foster recurrence of exposure, were selected, and included in the tasks. As mentioned before, the constructs' *usefulness* and *essentialness* were used to answer our question about the choice of words that were not so frequent.

4.2.2 #LancsBox tools findings in *Neighborhood* subset of COREL-GEO

In this section and the next, we will describe the functions of the concordancer tools, how they analyzed corpora contents, and we will mention the results they yielded to suggest future possibilities for a hands-on use in a school lab.

4.2.2.1 *Words* tool – word classes

First, the *Words* tool was used to generate the list of the most frequent words in the *Neighborhood* subset. The list is just a sample with the top words (Table 14). As it shows content and function words together, we needed to select only content words like nouns, adjectives, verbs, and adverbs (Appendix N). For the purposes of this investigation, the function words had to be disregarded as the content words were the main target language of this study. Although function words were not individually focused on as target language, some of them ended up being part of the study as many of them composed n-grams that the students were exposed to.

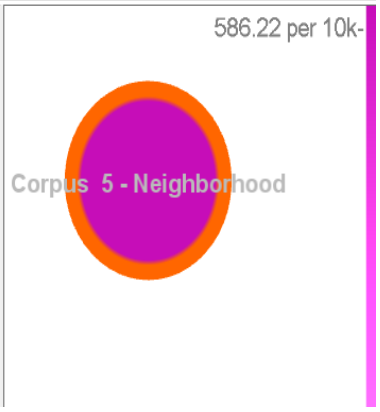
Table 14 – Most frequent words in subset *Neighborhood*

Corpora | Words: Corpus 5 - Neighborhood

Search

▼ Corpus | Corpus 5 - Neighborhood | ▼ Frequency | ▼ Dispersion | ▼ Type

Type	▼ Frequency: 01 - Freq	Dispersion: 01_CV
the	342.000000	0.462163
a	211.000000	0.757808
to	205.000000	0.678979
and	143.000000	0.716123
of	122.000000	0.650416
in	119.000000	0.593836
is	117.000000	0.934618
you	97.000000	1.269715
people	97.000000	1.150749
are	84.000000	0.879530
your	52.000000	1.846311
it	47.000000	1.487733
there	44.000000	1.478631
map	44.000000	1.614895
they	41.000000	1.228004



Source: #LancsBox - Tool: *Words*.

Second, the *Words* tool generated lists of content words filtered out in word classes (Tables 15 to18).

Table 15 – Most frequent nouns in the *Neighborhood* subset

▼ Corpus | Corpus 5 - Neighborhood | ▼ Frequency | ▼ Dispersion | ▼ Lemma

Lemma	▼ Frequency: 01 - Freq	Dispersion: 01_CV
people_n	96.000000	1.155172
city_n	61.000000	1.674136
place_n	54.000000	1.458141
map_n	47.000000	1.665468
community_n	47.000000	1.978356
school_n	35.000000	1.736022
there_other	31.000000	1.669198
home_n	26.000000	1.950122
world_n	21.000000	4.300135
building_n	21.000000	2.209308
area_n	20.000000	2.627697
find_v	19.000000	1.853134
library_n	19.000000	2.029869
question_n	18.000000	1.481799
street_n	17.000000	2.754022
neighborhood_n	16.000000	3.437658
photo_n	16.000000	1.639778
park_n	16.000000	2.198210
land_n	16.000000	3.516176
market_n	16.000000	3.553789
direction_n	15.000000	3.749130
house_n	15.000000	1.851981
population_n	14.000000	4.103200
town_n	14.000000	3.358024
police_n	13.000000	2.833325
country_n	13.000000	3.556306
farmer_n	12.000000	2.931712
thing_n	12.000000	2.302623
station_n	12.000000	3.967828
food_n	12.000000	4.095858
theater_n	12.000000	2.851192
language_n	11.000000	6.480741
service_n	11.000000	4.327246

Source: #LancsBox - Tool: *Words*.

Table 16 – Most frequent verbs in the *Neighborhood* subset

▼ Corpus Corpus 5 - Neighborhood ▼ Frequency ▼ Dispersion ▼ L

Lemma	Frequency: 01 - Freq	Dispersion: 01_CV
be_v	257.000000	0.495618
have_v	57.000000	0.948632
can_v	37.000000	1.520050
live_v	35.000000	1.715195
do_v	34.000000	1.633956
help_v	32.000000	2.523549
may_v	26.000000	2.329814
use_v	23.000000	2.011185
go_v	23.000000	2.279979
might_v	19.000000	3.268172
find_v	19.000000	1.853134
move_v	19.000000	3.837661
get_v	18.000000	2.063721
read_v	18.000000	1.349792
make_v	18.000000	2.036438
need_v	17.000000	2.290284
answer_v	16.000000	1.597273
give_v	15.000000	2.315234
come_v	15.000000	3.438425
want_v	14.000000	1.926907
take_v	12.000000	2.034406
grow_v	11.000000	4.002389
tell_v	11.000000	3.717114
would_v	10.000000	2.722362
keep_v	10.000000	2.858185
look_v	9.000000	2.253772
work_v	9.000000	2.903935
study_v	9.000000	2.370655
will_v	9.000000	3.502668
sell_v	9.000000	4.269962
visit_v	8.000000	3.368827
could_v	8.000000	3.412251

Source: #LancsBox – Tool: *Words*.

Table 17 – Most frequent adjectives in the *Neighborhood* subset

▼ Corpus Corpus 5 - Neighborhood ▼ Frequency ▼ Dispersion ▼ L

Lemma	Frequency: 01 - Freq	Dispersion: 01_CV
many_adj	35.000000	1.320839
rural_adj	17.000000	2.682700
different_adj	16.000000	1.853968
other_adj	15.000000	2.514234
more_adj	12.000000	3.357533
few_adj	11.000000	2.267644
large_adj	10.000000	4.034386
such_adj	9.000000	2.581147
urban_adj	9.000000	3.669938
good_adj	9.000000	3.879512
small_adj	9.000000	2.988021
big_adj	8.000000	2.556353
suburban_adj	6.000000	3.960460
public_adj	6.000000	3.034805
most_adj	6.000000	3.023783
natural_adj	6.000000	4.540614
new_adj	6.000000	2.829009
open_adj	5.000000	3.880316
busy_adj	5.000000	2.906156
safe_adj	5.000000	4.645469
local_adj	5.000000	3.772602
close_adj	5.000000	3.795491
official_adj	5.000000	6.480741
special_adj	4.000000	3.993565
near_adj	4.000000	3.850986
pet_adj	4.000000	6.480741
high_adj	4.000000	4.200514
easy_adj	3.000000	3.686901
short_adj	3.000000	4.697126
2nd_adj	3.000000	4.116360
great_adj	3.000000	4.849667
fast_adj	3.000000	4.784233

Source: #LancsBox – Tool: *Words*.

Table 18 – Most frequent adverbs in the *Neighborhood* subset

▼ Corpus Corpus 5 - Neighborhood ▼ Frequency ▼ Dispersion ▼ L

Lemma	▼ Frequency: 01 - Freq	Dispersion: 01_CV
then_adv	25.000000	1.131180
how_adv	22.000000	1.500205
where_adv	22.000000	2.004301
also_adv	17.000000	1.568237
not_adv	16.000000	1.679970
out_adv	15.000000	2.572923
when_adv	14.000000	1.825782
why_adv	13.000000	1.989796
up_adv	13.000000	2.206166
there_adv	13.000000	2.587732
here_adv	12.000000	3.163741
very_adv	8.000000	2.611174
more_adv	6.000000	3.300219
so_adv	6.000000	3.161966
together_adv	6.000000	3.005613
below_adv	5.000000	2.856532
too_adv	5.000000	4.216524
sometimes_adv	5.000000	2.936133
still_adv	4.000000	3.199796
close_adv	4.000000	3.767509
usually_adv	4.000000	3.589071
south_adv	4.000000	3.687625
just_adv	4.000000	3.169457
far_adv	3.000000	3.657377
all_adv	3.000000	3.833314
often_adv	3.000000	4.190150
north_adv	3.000000	3.725278
most_adv	3.000000	4.309134
apart_adv	3.000000	3.708253
home_adv	3.000000	4.072224
else_adv	3.000000	3.742013
down_adv	3.000000	3.741092

Source: #LancsBox – Tool: *Words*.

The process is very user-friendly: one should left-click on *Type* on the right at the top blue header above, change *Type* to *Lemma* by clicking on the arrow, and then on Apply. After that, right-click on the black bar, next to the word Type on the left above, and a pop-up window will open. Add: *_v, or *_n, or *_adj or *_adv, one at a time, to have the most frequent words of the different word classes. Click on Apply.

4.2.2.2 *N-grams* tool – word clusters

Another tool, *Ngrams,* was used to generate the word clusters. An n-gram is a contiguous sequence of *n* items that comes from a text or a corpus. Some of them, though frequent, may not be pedagogically relevant, so teachers should choose at their discretion those more meaningful to their class. The young learners are beginning to be exposed to English with a greater focus on the target language and the number of words in the clusters and their combinations may make a difference in their comprehension. In Table 19, the more frequent n-grams *answer the questions*, then *answer the* and *study the photo,* for example, are used for instructions which are quite often used in class and may not need to be highlighted for this study. Other 3-grams, such as, *is a lot* and *this map of a* are phrase fragments that would not be relevant to be taught. However, the prepositional phrases *in a community* and *on the map* as

well as the verb phrase *is a lot of* would be useful for young learners on many other occasions and with other topics as well. It has been put forward by many scholars that if prepositions and their collocates are taught in context from the very beginning of their exposure to young learners, they will be acquired in more appropriate combination with other lexis (Brown, 1973; Bolinger, 1976; Alexander, 1979; Carter, 1987; Wray, 2012). This is made possible with the work with concordance lines and the surrounding neighbors of the nodes. They are learned as fixed formulaic expressions and can denote higher degrees of fluency of the learner.

Table 19 – Most frequent 3- and 4-grams in the *Neighborhood* subset

▼ **Corpus** Corpus 5 - Neighborhood ▼ **Frequency** ▼ **Dispersion** ▼ T

Type	▼ Frequency: 01 - Freq	Dispers
answer the questions	18.000000	1.496661
then answer the	15.000000	1.710069
study the photo	13.000000	1.753243
this is a	13.000000	6.480741
read the text	13.000000	1.676720
the photo then	11.000000	1.874867
photo then answer	10.000000	1.943026
and study the	8.000000	2.265642
text and study	8.000000	2.265642
the text and	8.000000	2.265642
on the map	7.000000	2.825933
map of a	7.000000	2.334853
there is a	6.000000	2.841878
a lot of	6.000000	3.687286
the natural world	6.000000	4.541336
the united states	6.000000	6.004359
live in a	6.000000	3.020852
a map of	5.000000	3.180265
a farmers market	5.000000	3.993006
look at this	5.000000	2.843496
you want to	5.000000	3.340437
in the city	5.000000	4.372773
at a farmers	4.000000	4.169114
their home countries	4.000000	5.459973
of the natural	4.000000	4.541336
is a lot	4.000000	3.756763
in a community	4.000000	4.528115
you live in	4.000000	3.754058
a good citizen	4.000000	6.480741
the land in	4.000000	5.640306
lot of open	4.000000	4.129518
to get to	4.000000	3.687060

Source: #LancsBox - Tool: *Ngrams*.

▼ Corpus Corpus 5 - Neighborhood ▼ Frequency ▼ Dispersion ▼ T

Type	▼ Frequency: 01 - Freq	Dispers
then answer the questions	15.000000	1.719387
study the photo then	11.000000	1.877468
photo then answer the	10.000000	1.945319
the photo then answer	10.000000	1.945319
text and study the	8.000000	2.269430
read the text and	8.000000	2.269430
the text and study	8.000000	2.269430
and study the photo	7.000000	2.463162
to the united states	4.000000	6.480741
a lot of open	4.000000	4.130487
of the natural world	4.000000	4.541720
is a lot of	4.000000	3.756430
there is a lot	4.000000	3.756430
look at this map	4.000000	3.173215
at a farmers market	4.000000	4.182922
be used to make	3.000000	6.480741
at this map of	3.000000	3.681788
north west east south	3.000000	3.733610
this map of a	3.000000	3.681788
public transportation to get	3.000000	3.718704
read the text study	3.000000	3.658853
lot of open space	3.000000	5.304817
the text study the	3.000000	3.658853
text study the photo	3.000000	3.658853
in their home countries	3.000000	6.480741
a rural community has	3.000000	4.856845
how do you know	3.000000	4.167569
transportation to get to	3.000000	3.718704
answer the questions urban	3.000000	4.040550
you are going to	2.000000	6.480741
where many people live	2.000000	5.048539
what kind of community	2.000000	6.480741

Source: #LancsBox - Tool: *Ngrams*.

4.2.3 Choice of vocabulary in *Neighborhood* (T1)

Two software tools were used to determine not only the frequency and dispersion of words in the *Neighborhood* subset but also to establish the number of KWIC occurrences in concordance lines distributed in the 43 texts. These numbers were used as an indicator of the degree of usefulness of the KWICs to be handled by learners. Among the lexis selected for this investigation, some words like *neighborhood*, *community*, *supermarket* vs *market*, *suburb* vs *suburbs* in their concordance lines are illustrated below. One useful example is the comparison between *market* and *supermarket* selected to clarify cultural uses of the two words and to justify their inclusion in the tasks.

4.2.3.1 *KWIC* and *Words* tools

With the tool *KWIC* it is also possible to see inside the corpus and identify the files and/or texts where the target word (KWIC) is present shown on the left of Figure 22. The first

one, *neighborhood*, has 14 occurrences in 6 texts, which indicates its high potential for inclusion in the tasks and tests. One suggestion for the future work is that learners receive many concordance lines, as in Figure 22, to compare and contrast them to notice the patterns in the sentences and also the neighbors of the node in red.

Figure 22 – KWIC *neighborhood*

Search neighborhood **Occurrences** 14 (24.00) Texts 6/43

Index	File	Left	Node	
1	Geo_2ndG_Ir	to keep you, your family, and your	neighborhood	safe. They are a very important part
2	Geography_1	A	Neighborhood	Map A neighborhood map often uses pictures
3	Geography_1	A Neighborhood Map A	neighborhood	map often uses pictures of buildings as
4	Geography_1	Map Your	Neighborhood	There are all kinds of neighborhoods. Some
5	Geography_1	very few buildings. A map of your	neighborhood	can help you understand the things you
6	Geography_1	there. It can help you describe your	neighborhood	to a friend. Imagine the square in
7	Geography_1	the home you live in. Draw your	neighborhood	around where you live. Include squares for
8	Geography_3	else. Look at this map of a	neighborhood.	HIGH school Garden center Gas station Elementary
9	Geography_3	Map Your	Neighborhood	Maps are tools that help you understand
10	Geography_3	away. Create a map of your own	neighborhood	in the grid below. Give your map
11	Video_Geo_M	this video we will learn about our	neighborhood	neighborhood means places near us the area
12	Video_Geo_M	video we will learn about our neighborhood	neighborhood	means places near us the area around
13	Video_Geo_M	area around our house is called our	neighborhood	the houses built close to each other
14	Video_Geo_M	each other make up our neighbors our	neighborhood	has many services now we will go

Source: #LancsBox – Tool: *KWIC*.

Table 20 displays the distribution of the target word in texts, which is another area to be explored with learners. From this information they can select those files with more occurrences to investigate and discover the language patterns.

Table 20 – Noun: *neighborhood* – distribution in texts and relative frequency per 10k

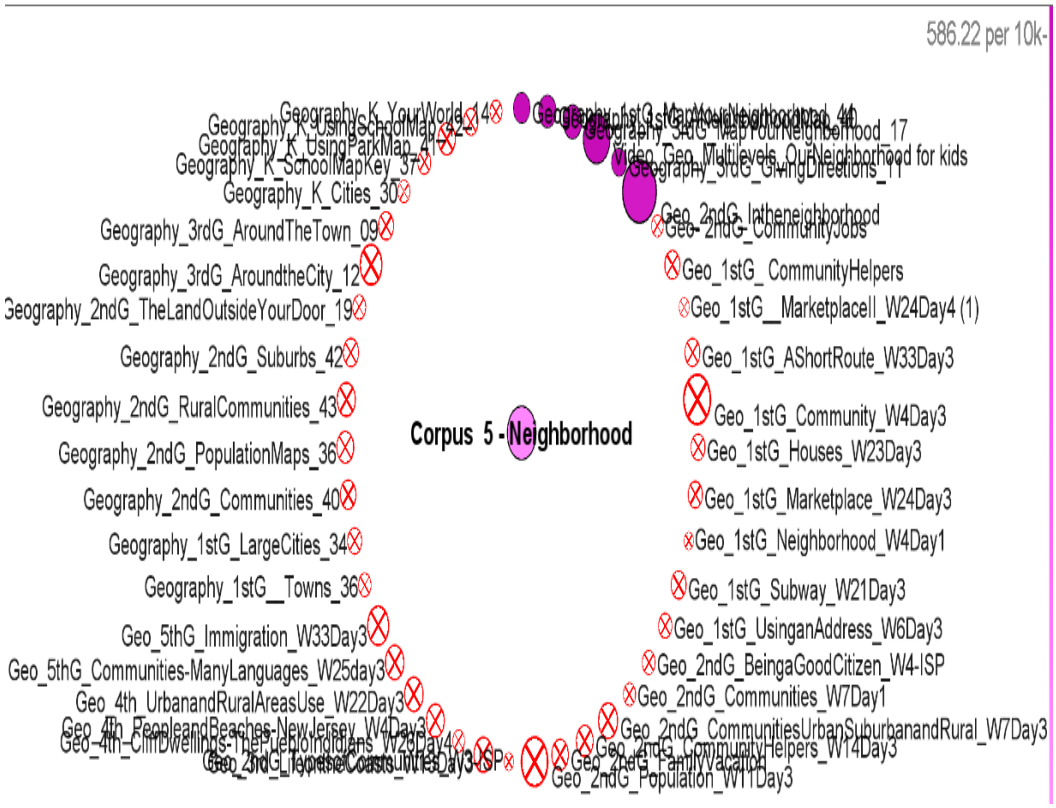
Corpus 5 - Neighborhood: neighborhood_n

File	Tokens	Frequency	▼ Relative f
Geography_1stG_MapYourNeig...	122	6	491.80328
Geography_1stG_ANeighborho...	137	2	145.9854
Geography_3rdG_MapYourNeig...	150	2	133.33334
Video_Geo_Multilevels_OurNei...	318	4	125.786156
Geography_3rdG_GivingDirectio...	102	1	98.03922
Geo_2ndG_Intheneighborhood.txt	512	1	19.53125

Source: #LancsBox – Tool: *Words*.

In the future, having general information and communication technolgy (ICT) skills and with the hands-on use of the software, the teacher can call learners' attention to how the target language is distributed in the subset of corpus, resorting also to the distribution chart (Figure 23) below. The chart is visually attractive and will impact on the traditional approach to tables and figures. Learners can explore those files which show bigger and darker dots at the top on the right where there is more concentration of the KWIC being investigated.

Figure 23 – Distribution of *neighborhood* inside the corpus and location in the files



Source: #LancsBox – Tool: *Words.*

A second example is the word *community* in 5th place (Table 15) in the list of most frequent words. Its 36 hits (Figure 24) and their concordance lines in 11 texts can offer a considerable recurrence of exposure of the word to learners. In this case, the teacher selected the lines which could provide learners with immediate comprehension:

- for the 4th graders: *What is a community? What kind of community do you live in? Types of community service*, *A rural community has a lot of open space*, for example; and

- for more advanced 5th and 6th graders: *People help out in the community*, *What is community Service?*, *A community is made up of places to go*, *A suburban community is a smaller town near a larger city.*

Figure 24 – KWIC: *community*

Search community **Occurrences** 36 (61.71) Texts 11/43

Index	File	Left	Node	
1	Geo- 2ndG_C	DIFFERENT JOBS Many different people in your	community	work to provide goods and services to
2	Geo- 2ndG_C	others. When people work, they help the	community	and earn an income. For example, a
3	Geo_1stG_ C	word bank below to fill in the	community	helper sentences. Then draw a line connecting
4	Geo_1stG_C	A	Community	What is a community? It is a
5	Geo_1stG_C	A Community What is a	community?	It is a place where people live.
6	Geo_1stG_C	there. There are special places in a	community.	There might be a zoo. There might
7	Geo_1stG_C	zoo. There might be an airport. A	community	is made up of places to go,
8	Geo_1stG_C	Name two places people go in a	community.	Is a museum part of a community?
9	Geo_1stG_C	community. Is a museum part of a	community?	How do you know? A rural community
10	Geo_1stG_C	community? How do you know? A rural	community	has a lot of open space. It
11	Geo_1stG_C	market is a special place in a	community.	Study the photo. Answer the questions. Write
12	Geo_1stG_N	Where could someone buy dessert in this	community?	How many gas stations are in this
13	Geo_1stG_N	How many gas stations are in this	community?	
14	Geo_2ndG_C	over very large areas. What type of	community	does the cow live in? How do
15	Geo_2ndG_C	How do you know? What type of	community	is behind the cow? What might happen
16	Geo_2ndG_C	close to many people, what kind of	community	do you live in? If there is
17	Geo_2ndG_C	lot of open space, what kind of	community	do you live in? Which two types
18	Geo_2ndG_C	study the photo. Then, answer the questions.	Community	Helpers People help out in the community.
19	Geo_2ndG_C	Community Helpers People help out in the	community.	Bus drivers take people where they need
20	Geo_2ndG_C	else could you help out in your	community?	
21	Geo_2ndG_Ir	business Tax Collection Library Housing What is	Community	service? Community service is volunteering to help
22	Geo_2ndG_Ir	Collection Library Housing What is Community service?	Community	service is volunteering to help those in
23	Geo_2ndG_Ir	is volunteering to help those in your	community.	It might be something that is done
24	Geo_2ndG_Ir	done once or on a regular basis,	Community	service is often referred to as `giving
25	Geo_2ndG_Ir	done by an individual or an organization.	Community	service is giving your time without being
26	Geo_2ndG_Ir	fortunate or to help clean up your	community.	Types of community service: Soup kitchen Homeless
27	Geo_2ndG_Ir	help clean up your community. Types of	community	service: Soup kitchen Homeless shelter Mentoring Tutoring
28	Geo_2ndG_Ir	a line from a situation to the	community	service or person: You need to research
29	Geo_2ndG_Ir	are a very important part of your	community.	Find out the following information about your
30	Geo_2ndG_T	city urban buildings rural country An urban	community	is filled with people, buildings, and different
31	Geo_2ndG_T	and different types of transportation. A suburban	community	is a smaller town near a larger
32	Geo_2ndG_T	town near a larger city. A rural	community	has open space and few homes in
33	Geography_2	Communities A	community	is a group of people or animals
34	Geography_2	suburban, or rural. I live in a	community	
35	Geography_2	Rural Communities A rural	community	has fewer people than a suburb. There
36	Geography_2	Look at the map of a rural	community	below. Then complete the activities. NORTH West

Source: #LancsBox – Tool: *KWIC.*

Table 21 – Noun: *community* – distribution in texts and relative frequency per 10k

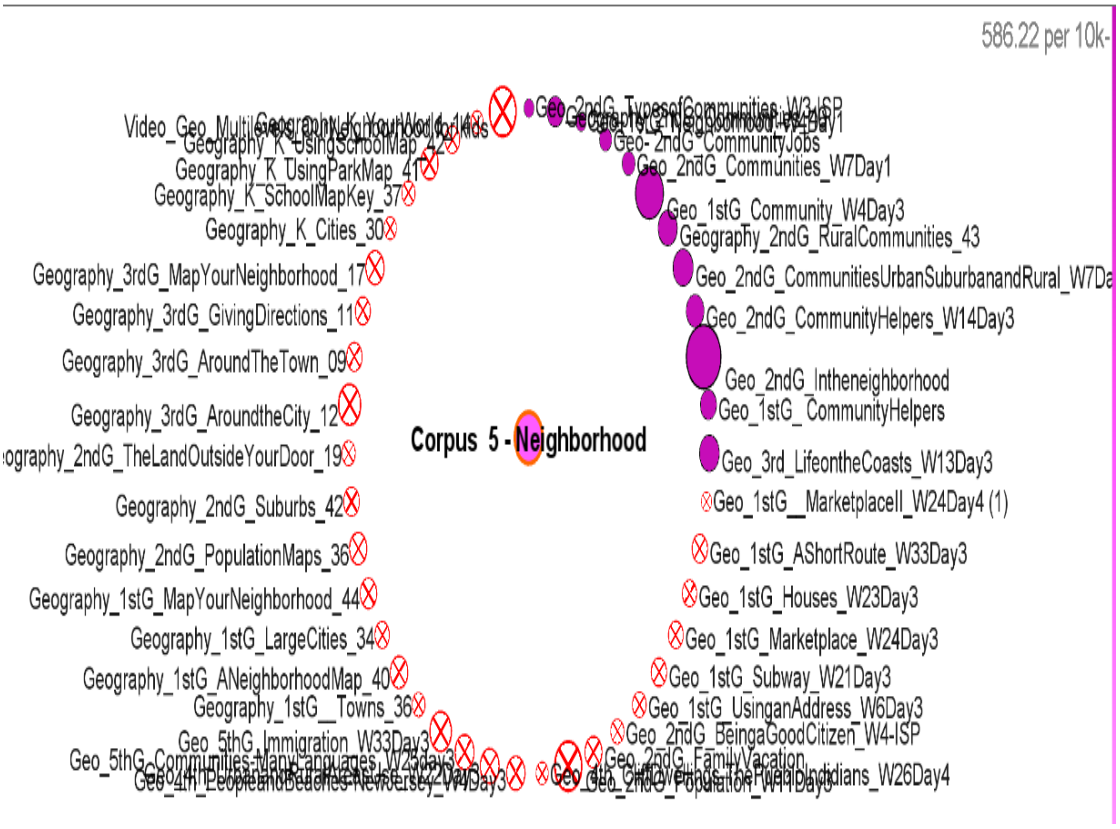
Corpus 5 - Neighborhood: community_n

File	Tokens	Frequency	▼ Relative frequency per 10k
Geo_2ndG_TypesofCommunitie...	42	3	714.28577
Geography_2ndG_Communities...	116	7	603.44824
Geo_1stG_Neighborhood_W4D...	41	2	487.80487
Geo- 2ndG_CommunityJobs.txt	64	2	312.5
Geo_2ndG_Communities_W7D...	70	2	285.7143
Geo_1stG_Community_W4Day3...	335	9	268.65674
Geography_2ndG_RuralCommu...	155	4	258.0645
Geo_2ndG_CommunitiesUrban...	171	4	233.91814
Geo_2ndG_CommunityHelpers_...	133	3	225.5639
Geo_2ndG_Intheneighborhood.txt	512	9	175.78125
Geo_1stG_ CommunityHelpers....	110	1	90.90909
Geo_3rd_LifeontheCoasts_W13...	166	1	60.240963

Source: #LancsBox – Tool: Words.

The internal distribution of the corpora shown in Table 21, and displayed in Figure 25 below, can be very effective for consolidation work and future retention of the target word.

Figure 25 – Distribution of *community* inside the corpus and location in the files



Source: #LancsBox – Tool: *Words.*

The next lexis, *market*, is in the 20th place in Table 15 but was considered to be very relevant to the students' real life. In Brazil, people living in big cities, which is the case of the study's participants, seem to use the word *supermarket* more than *market*, leading the teacher to want to contrast the use of these words. Below it is possible to identify only 6 occurrences of *supermarke*t in 5 texts (Figure 26) while there are 16 occurrences of *market* also in 5 texts (Figure 27).

Figure 26 – KWIC: *supermarket*

Search supermarket **Occurrences** 6 (10.28) Texts 5/43

Index	File	Left	Node	
1	Geography_1	for buildings like a school or a	supermarket.	Your home What kind of building is
2	Geography_2	store gas station restaurant doctor's office pharmacy	supermarket	playground
3	Geography_3	school Elementary school Gas station Garden center	Supermarket	Movie theater Drugstore Hospital Index Elementary school
4	Geography_3	Elementary school Movie theater Hospital Garden center	Supermarket	Gas station Drugstore High school
5	Geography_3	school Garden center Gas station Elementary school	Supermarket	Movie theater Drugstore Hospital Using this map,
6	Video_Geo_M	go through these services this is the	supermarket	it has many shops we go to

Source: #LancsBox – Tool: *KWIC.*

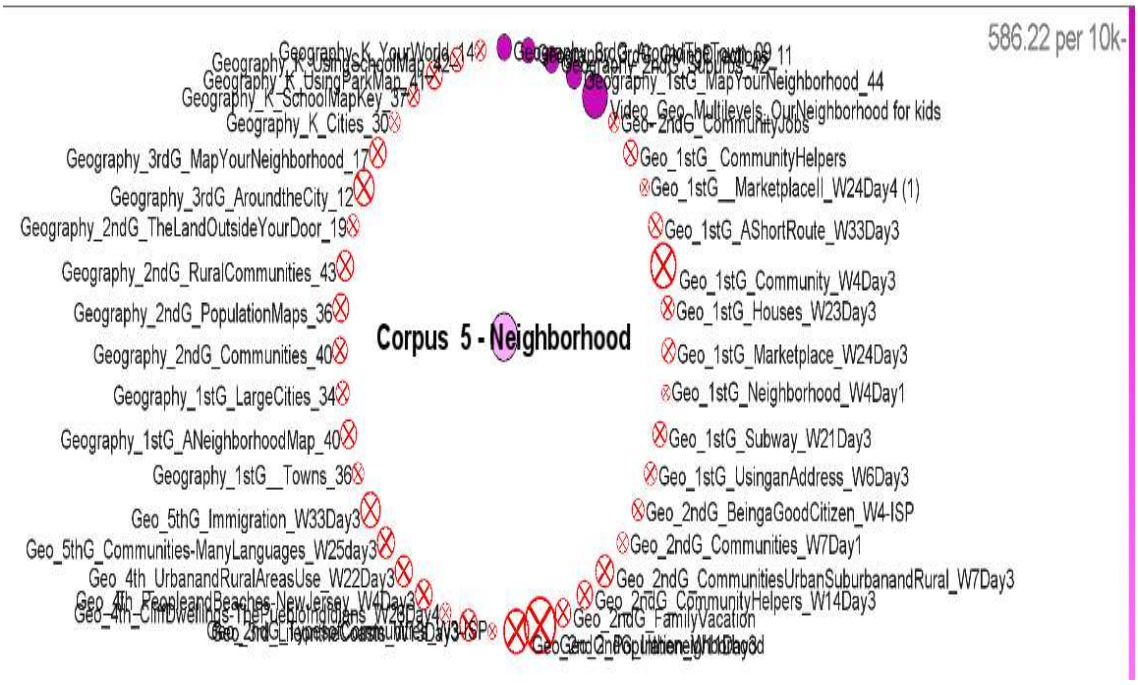
Table 22 – Noun: *supermarket* – distribution in texts and relative frequency per 10k

Corpus 5 - Neighborhood: supermarket_n

File	Tokens	Frequency	▼ Relative frequency per 10k
Geography_3rdG_AroundTheTo...	108	2	185.18518
Geography_3rdG_GivingDirectio...	102	1	98.03922
Geography_2ndG_Suburbs_42.txt	110	1	90.90909
Geography_1stG_MapYourNeig...	122	1	81.96721
Video_Geo_Multilevels_OurNei...	318	1	31.446539

Source: #LancsBox – Tool: *Words.*

Figure 27 – Distribution of *supermarket* inside the corpus and location in the files



Source: #LancsBox – Tool: *Words.*

The dark dots in Figure 27 correspond to the information in Table 22 which makes it visually clear to learners where the most recurrent target language can be found. In the case of *supermarket*, there are 3 files which could be investigated. Conversely, *market* has 16 occurrences in the same number of texts (Table 23), clearly illustrated in Figure 28. An activity which may trigger learners' curiosity is to investigate the texts with the darker dots in Figures 27 and 29 and compare the usage – collocations and colligations – of findings. They could play the role of language detectives (Johns, 1991), a key pillar in the DDL approach.

Figure 28 – KWIC: *market*

Search market **Occurrences** 16 (27.43) Texts 5/43

Index	File	Left	Node	
1	Geo_1stG__	shows the goods sold at a farmers'	market	one morning. Study the graph. Then, answer
2	Geo_1stG__	the graph. Then, answer the questions. Farmers'	Market	Morning Sales Sales honey nuts vegetables fruits
3	Geo_1stG_(	wheat be used to make? A farmers'	market	is a special place in a community.
4	Geo_1stG_(	Write about what happens at a farmers'	market.	List things people can buy at a
5	Geo_1stG_(	things people can buy at a farmers'	market.	Food from Around the World You can
6	Geo_1stG_M	photo. Then, answer the questions. Pike Place	Market	Pike Place is in Seattle, Washington. Goods
7	Geo_1stG_M	booths to show their art. Pike Place	Market	is an open-air market. But it has
8	Geo_1stG_M	art. Pike Place Market is an open-air	market.	But it has a roof. So people
9	Geo_1stG_M	people come to shop-rain or shine. FARMERS	MARKET	Where do the goods at a farmers'
10	Geo_1stG_M	Where do the goods at a farmers'	market	come from? Why do artists come to
11	Geo_1stG_M	Why do artists come to the farmers'	market?	
12	Geography_	the A to meet the Arch St.	Market	St. Chestnut St. Harbor Area Walnut St.
13	Geography_	triangle at the park. The Reading Terminal	market	is in It lies between Arch and
14	Geography_	is in It lies between Arch and	Market	streets to the north and south, and
15	Geography_	a shopping bag where the Reading Terminal	market	is located.
16	Video_Geo_	has many shops we go to the	market	to buy things this is a hospital

Source: #LancsBox – Tool: *KWIC.*

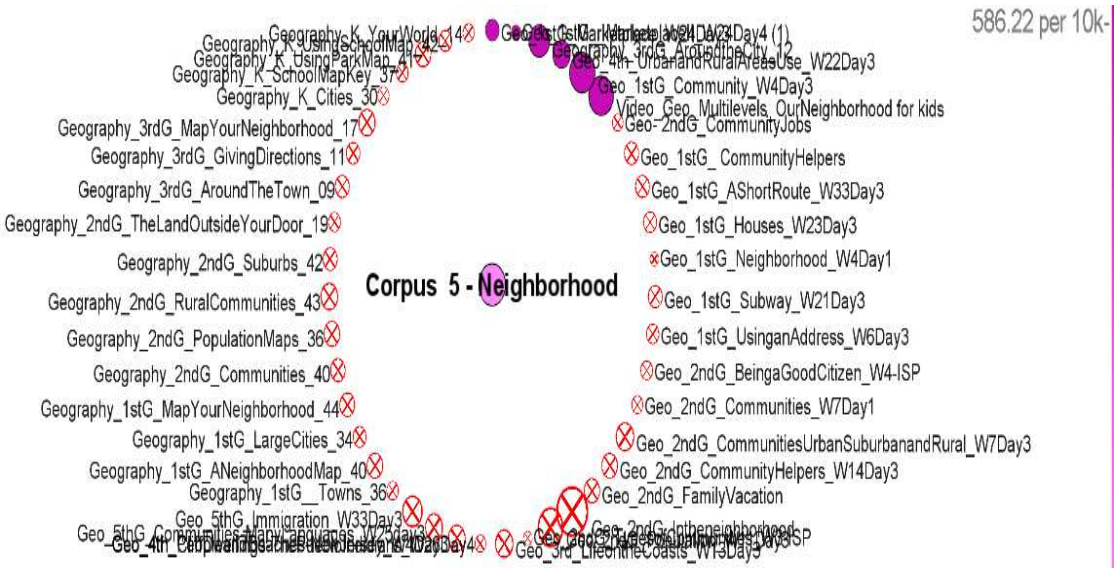
Table 23 – Noun: *market* – distribution in texts and relative frequency per 10k

Corpus 5 - Neighborhood: market_n

File	Tokens	Frequency	▼ Relative frequency per 10k
Geo_1stG_Marketplace_W24Da...	98	6	612.24493
Geo_1stG__MarketplaceII_W24...	47	2	425.5319
Geography_3rdG_AroundtheCity...	220	4	181.81818
Geo_4th_UrbanandRuralAreas...	161	1	62.1118
Geo_1stG_Community_W4Day...	335	2	59.701492
Video_Geo_Multilevels_OurNei...	318	1	31.446539

Source: #LancsBox – Tool: *Words.*

Figure 29 – Distribution of *market* inside the corpus and location in the files



Source: #LancsBox – Tool*: Words.*

A further comparison of KWICs was also carried out, this time with the words *suburb* (Figure 30) and *suburbs* (Figure 32). It is also possible to recognize that *suburb*, 39th place in

the frequency list of subset *Neighborhood*, has just 3 occurrences in 2 texts, while *suburbs*, less frequent among learners in my country, has 7 occurrences in the 2 texts as well. *Suburb* is a cognate to *subúrbio* (singular noun) in Portuguese, but *suburb*s would not be used unless one would be referring to a plural noun as this is the correct grammatical usage for plural nouns.

Figure 30 – KWIC: *suburb*

Search suburb **Occurrences** 3 (5.14) Texts 2/43

Index	File	Left	Node	
3	Geography_2	buildings to complete the map of a	suburb	given below. You may take ideas of
2	Geography_2	smaller and not so busy. In a	suburb,	homes are still close together. People live
1	Geography_2	rural community has fewer people than a	suburb.	There is a lot of open land.

Source: #LancsBox – Tool: *KWIC.*

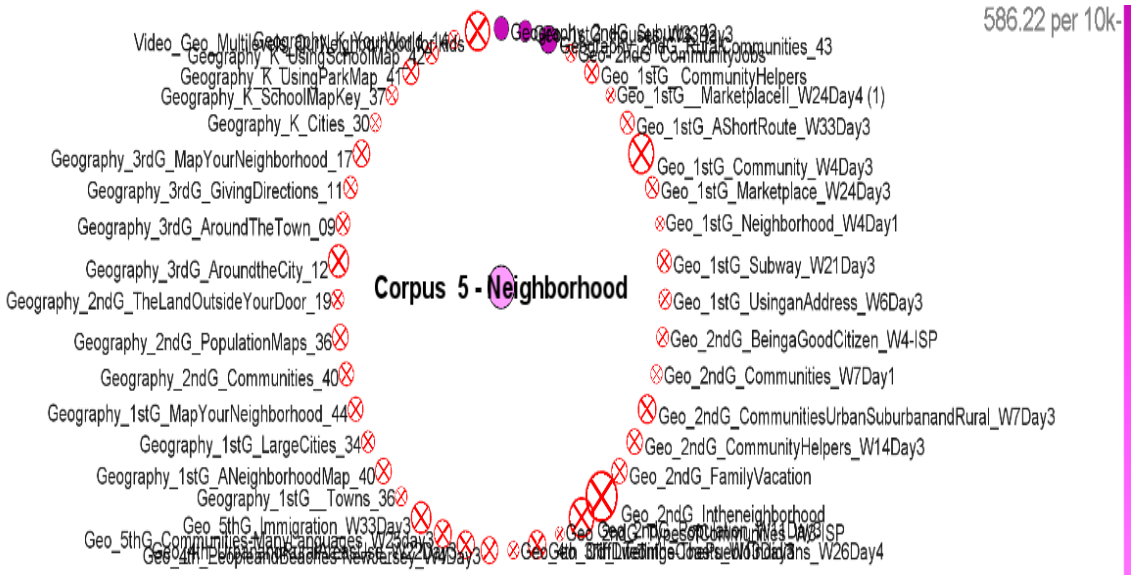
Table 24 – Noun: *suburb* – distribution in texts and relative frequency per 10k

Corpus 5 - Neighborhood: suburb_n

File	Tokens	Frequency	Relative frequency per 10k
Geography_2ndG_Suburbs_42.txt	110	5	454.54547
Geo_1stG_Houses_W23Day3.txt	93	3	322.58063
Geography_2ndG_RuralComm...	155	1	64.51613

Source: #LancsBox – Tool*: Words.*

Figure 31 – Distribution of *suburb* inside the corpus and location in the files



Source: #LancsBox – Tool: *Words.*

Figure 32 – KWIC: *suburbs*

Search suburbs **Occurrences** 7 (12.00) Texts 2/43

Index	File	Left	Node	
1	Geo_1stG_	Then, answer the questions. Rows of Homes	Suburbs	are often found outside cities. They are
2	Geo_1stG_	and backyards. People can work in the	suburbs.	But many people still work in the
3	Geo_1stG_	is not too long. How are the	suburbs	different from a city? How can people
4	Geo_1stG_	city? How can people live in the	suburbs	and work in the city?
5	Geography_		Suburbs	Suburbs are near cities, but are smaller
6	Geography_	Suburbs	Suburbs	are near cities, but are smaller and
7	Geography_	with a yard, or in an apartment.	Suburbs	also have schools, libraries, businesses, and parks.

Source: #LancsBox – Tool: *KWIC.*

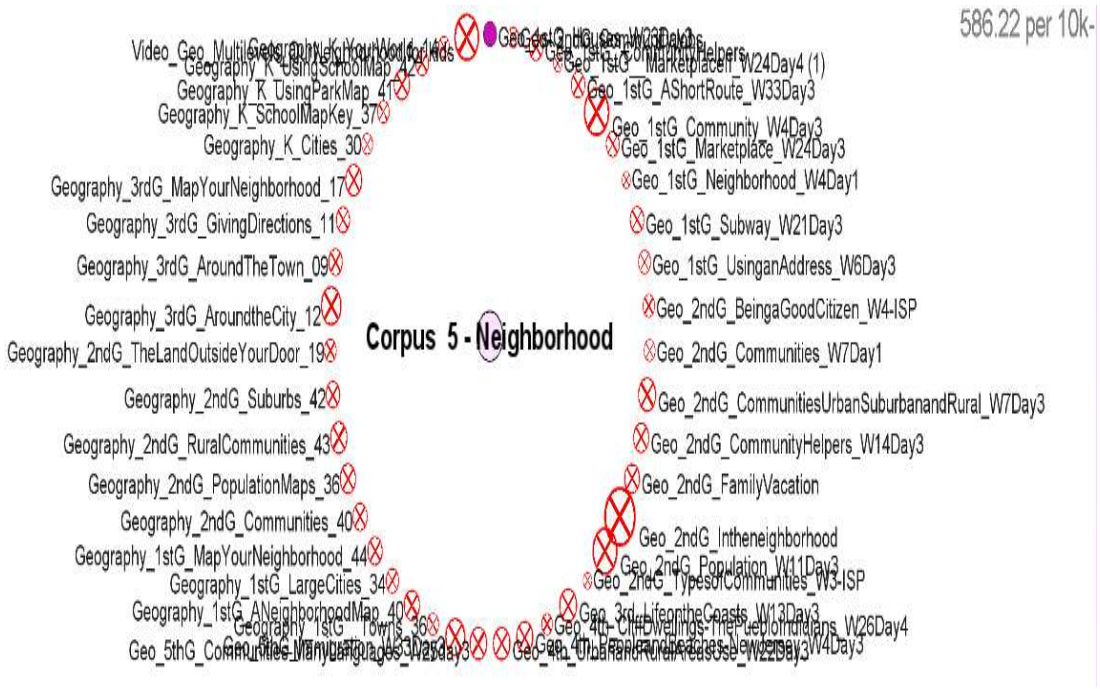
Table 25 – Noun: *suburbs* – distribution in texts and relative frequency per 10k

Corpus 5 - Neighborhood: suburbs_n

File	Tokens	Frequency	Relative frequency per 10k
Geo_1stG_Houses_W23Day3.txt	93	1	107.52688

Source: #LancsBox- Tool: *Words.*

Figure 33 – Distribution of *suburbs* inside the corpus and location in the files



Source: #LancsBox – Tool: *Words.*

The same procedures with the tools KWIC and *Words* above were used with the other word classes such as adjectives, verbs, and adverbs (Appendix L) for the identification of more words to be worked with using the DDL approach.

4.2.3.2 *Text* tool

To complement and clarify learners' doubts during the lessons, a fourth tool was used: *Text.* When difficulties arose, and before learners produced personalized sentences with the KWICs, the teacher resorted to the original context to clarify comprehension. One such example is one of the original contexts for *neighborhood* selected below. The tool can pop-up any original context in the corpus of the word selected.

Figure 34 – Short extract of *neighborhood* context

Map Your Neighborhood
There are all kinds of neighborhoods. Some neighborhoods have many different kinds of buildings.
Others may have very few buildings. A map of your neighborhood can help you understand the things you find there.
It can help you describe your neighborhood to a friend.
Imagine the square in the middle of the box below is the home you live in. Draw your neighborhood around where you live.
Include squares for buildings like a school or a supermarket.
Your home
What kind of building is your home Color the middle square the same color as your real home.
On your map, which two buildings are closest to your home? Color those two squares the same color as the real buildings.

Source: #LancsBox – Tool: *Text.*

Following McCarthy's (2004) and O'Keeffe (2021) recommendation that teachers should mediate data to find the clearest and best examples to use from the corpus, examples of concordance lines were copied and cut out in strips of paper to be dealt with by students in the classroom (Figure 35). The lines were copied and enlarged to make it easier for the youngsters to observe the KWICs and the surrounding contexts. The words and clusters in bold were emphasized by the teacher during classwork.

Figure 35 – Samples of the assorted concordance lines selected

There are many different types of **map.** This is because we use different **maps** for different reasons.

If you were **going on a hike** in a **nature park**, you would need a map of the **park.**

If you wanted to know the way around your **neighborhood**, you would need a map of your neighborhood.

find the best road to your **friend's house**

find your way around a **nature park**

Cities are busy places! They have many **buildings**, including businesses and schools.

City block with **homes** and **stores**

Many of the people you know probably live in your **neighborhood**.

There is a **library** nearby

Source: #LancsBox.

4.2.3.3 *GraphColl* tool

The last tool used, *GraphColl,* is the one whose results are the most attractive to learners. Because its language analysis chart presents a colorful visual impact, it draws learners' attention to the KWIC collocates. It lends itself to assorted activities to be designed by teachers. The tool lines and dots can enhance learners' observation and noticing of the selected KWICs nodes of the concordance lines. The collocates can be seen in different colors magnifying the focus. The resulting graph is especially beneficial when working with colligations in the classroom and can be even more effective with advanced learners.

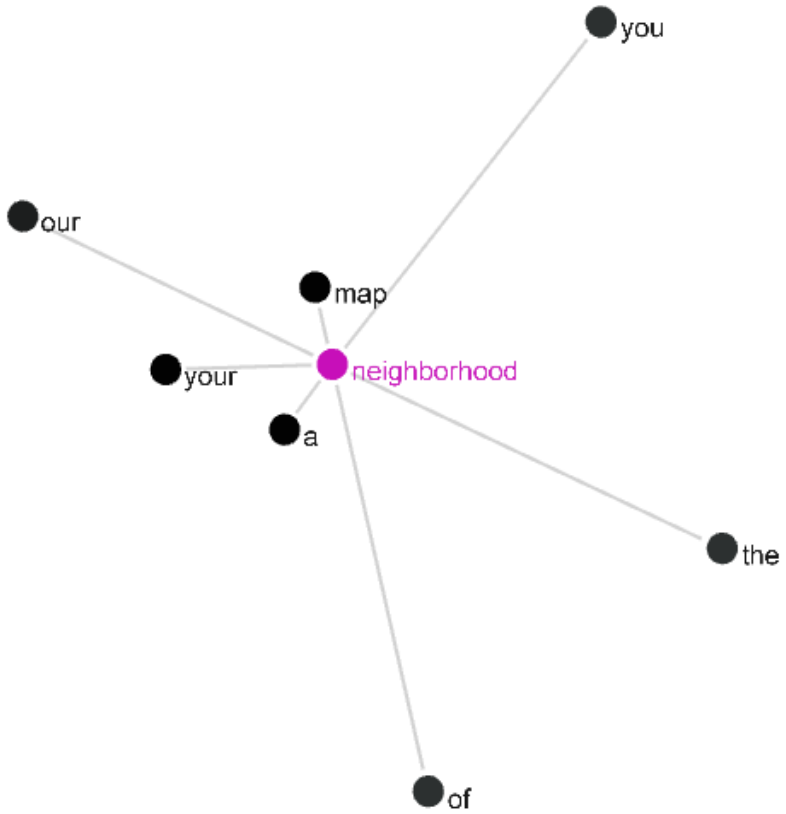
The findings around the node *neighborhood* (Chart 1) show three dimensions:

1) the strength of the collocation, indicated by the statistical measure MI score. The closer the collocate is to the node, the more strongly associated it is;

2) the frequency, indicated by the intensity of the color of the collocate dot. The darker the color, the more frequent it is; and

3) the position of the collocate around the node, either left or right, shows the actual position in the sentence.

Chart 1 – *Neighborhood* and its 7 collocates



Source: #LancsBox – Tool: *GraphColl.*

Table 26 – Collocates and its positions and frequency around the node *neighborhood*

neighborhood

Freq: 14 - Collocates: 7

Index	Status	Position	Collocate	▼ Stat	Freq (coll.)	Freq (corpus)
1	○	L	a	9.0	9	211
2	○	L	map	9.0	9	44
3	○	L	your	8.0	8	52
4	○	L	our	6.0	6	9
5	○	L	of	5.0	5	122
6	○	R	the	5.0	5	342
7	○	L	you	5.0	5	97

Source: #LancsBox – Tool: *GraphColl.*

If one wants to see any of the above collocates in their actual context in a concordance line, they can right click on the collocate in the graph (Chart 1) and a pop-up will show the node in red and the collocate in blue (Figures 36-37):

Figure 36 – Pop-up showing the position of collocate *map*

Search neighborhood **Occurrences** 7/14 (12.00) Texts 4/43 **▼ Corpus** Corpus 5 - Neighborhood **▼ Con**

Index	File	Left	Node	Right
2	Geography_1s	A	Neighborhood	Map A neighborhood map often uses pictures
3	Geography_1s	A Neighborhood Map A	neighborhood	map often uses pictures of buildings as
4	Geography_1s	Map Your	Neighborhood	There are all kinds of neighborhoods. Some
5	Geography_1s	very few buildings. A map of your	neighborhood	can help you understand the things you
8	Geography_3r	else. Look at this map of a	neighborhood.	HIGH school Garden center Gas station Elementary
9	Geography_3r	Map Your	Neighborhood	Maps are tools that help you understand
10	Geography_3r	away. Create a map of your own	neighborhood	in the grid below. Give your map

Source: #LancsBox – Tool: *GraphColl.*

Figure 37 – Pop-up showing the position of collocate *your*

KWIC: neighborhood > your ×

Search neighborhood **Occurrences** 7/14 (12.00) Texts 3/43 **▼ Corpus** Corpus 5 - Neighborhood **▼ Context** 7 **▼ Display T**

Index	File	Left	Node	Right
1	Geo_2ndG_Ir	to keep you, your family, and your	neighborhood	safe. They are a very important part
4	Geography_1	Map Your	Neighborhood	There are all kinds of neighborhoods. Some
5	Geography_1	very few buildings. A map of your	neighborhood	can help you understand the things you
6	Geography_1	there. It can help you describe your	neighborhood	to a friend. Imagine the square in
7	Geography_1	the home you live in. Draw your	neighborhood	around where you live. Include squares for
9	Geography_3	Map Your	Neighborhood	Maps are tools that help you understand
10	Geography_3	away. Create a map of your own	neighborhood	in the grid below. Give your map

Source: #LancsBox – Tool: *GraphColl.*

4.2.3.4 *N-grams* tool

Lastly, the *Ngrams* tool was used for the identification of the most frequent 3 and 4-grams. The teacher selected the ones which were among the most frequent and would be a

natural choice by learners when doing the activities. Figure 37 shows examples of clusters which were used by learners in their production in the posttests: *on a map, find your way* and *find your way around, different kinds of, need a map, live in a*, *in your neighborhood*, *find the best road to, there are many different types.* Others included can be seen in Appendix N.

Figure 38 – Samples of 3-grams used in tests and class activities

and you want
you are going
you are at
pond draw a
on a map
find your way
at the picture

your way around
imagine you are
from one place
can help you
if you keep
different kinds of
need a map

south east and
a map can
answer the questions
where something is
transportation to get
use the compass
live in a
in your neighborhood

Source: #LancsBox – Tool: Ngrams.

4.2.4 #LancsBox tools findings in *Animals* subset of COREL-SCI

As mentioned in the previous subsection 4.2.2, the *Words* tool was used to generate the list of the most frequent words in the *Animals* subset. The list is just a sample with the top words (Table 27).

Table 27 – Most frequent words in subset *Animals*

▼ Corpus Corpus 4 - Animals ▼ Frequency ▼ Dispersion ▼ Type

Type	▼ Frequency: 01 - Freq	Dispersion: 01_CV
the	445.000000	0.696336
are	322.000000	0.917598
and	240.000000	0.732283
a	238.000000	0.838505
of	231.000000	0.876134
to	168.000000	0.780823
they	166.000000	1.168842
in	163.000000	0.981534
that	150.000000	0.997775
have	141.000000	0.994414
animals	133.000000	1.433653
is	131.000000	1.150265
their	110.000000	1.280485
or	93.000000	1.816874
for	71.000000	1.567926
live	65.000000	1.840722
on	57.000000	1.476234
all	55.000000	2.062761
can	55.000000	1.673178
insects	54.000000	2.323662
some	50.000000	2.082452
it	50.000000	1.514425
called	46.000000	2.073590
animal	45.000000	1.822897
species	42.000000	2.794986
with	41.000000	2.078129
water	41.000000	2.569122
fish	40.000000	3.312953
mammals	40.000000	3.304764
what	40.000000	2.112426
do	40.000000	1.899661
like	40.000000	1.809024

Source: #LancsBox - Tool: *Words*.

Then, to generate just the content words in their individual classes, the tool was used again, adjusted to the demands. The findings are illustrated in Tables 28 – 32.

4.2.4.1 *Words* tool – word classes

Tables 28 – 31 show the content words and the 3- or 4-gram language clusters in the *Animals* subset. The findings also indicate the function words as the most frequent to the detriment of the content words. Once more the software tool had to be adjusted to yield nouns, verbs, adjectives, and adverbs as explained in subsection 4.1.2. The full list of results can be seen in Appendix O.

Table 28 – Most frequent nouns in the *Animals* subset

▼ Corpus Corpus 4 - Animals ▼ Frequency ▼ Dispersion ▼ Lemma

Lemma	Frequency: 01 - Freq	Dispersion: 01_CV
animal_n	165.000000	1.131762
insect_n	61.000000	2.611211
mammal_n	48.000000	3.500620
bird_n	44.000000	3.173543
fish_n	43.000000	3.338003
water_n	41.000000	2.527510
type_n	39.000000	2.900862
body_n	38.000000	2.121597
group_n	38.000000	2.817308
snake_n	34.000000	4.529454
wing_n	34.000000	2.494012
leg_n	33.000000	2.589348
backbone_n	33.000000	2.710849
specie_n	32.000000	3.055625
part_n	30.000000	2.181851
cat_n	30.000000	2.716561
vertebrate_n	28.000000	4.297410
mollusk_n	28.000000	5.753066
egg_n	26.000000	2.808295
spider_n	25.000000	3.338940
baby_n	25.000000	3.557181
reptile_n	24.000000	4.608601
kind_n	23.000000	2.996490
lion_n	21.000000	3.673580
people_n	21.000000	3.341577
turtle_n	21.000000	3.544087
question_n	19.000000	2.456059
frog_n	19.000000	5.038150
night_n	18.000000	6.619612
gill_n	17.000000	6.001531
world_n	17.000000	2.740542
day_n	17.000000	4.531817

Source: #LancsBox - Tool: Words

Table 29 – Most frequent verbs in *Animals* subset

▼ Corpus Corpus 4 - Animals ▼ Frequency ▼ Dispersion ▼ Lemma

Lemma	Frequency: 01 - Freq	Dispersion: 01_CV
be_v	494.000000	0.619691
have_v	172.000000	0.927506
live_v	77.000000	1.664547
can_v	61.000000	1.578228
do_v	55.000000	1.509275
call_v	48.000000	2.046402
use_v	35.000000	2.171928
fly_v	32.000000	3.024409
know_v	27.000000	1.983511
look_v	25.000000	2.479805
help_v	23.000000	3.111343
make_v	22.000000	2.159102
find_v	22.000000	2.552927
eat_v	20.000000	3.167151
protect_v	15.000000	3.571161
lay_v	15.000000	3.708034
breathe_v	15.000000	5.001170
relate_v	15.000000	3.064964
keep_v	14.000000	3.012413
get_v	13.000000	3.748312
move_v	13.000000	3.602592
belong_v	12.000000	3.566554
see_v	12.000000	3.437997
read_v	11.000000	2.735985
grow_v	10.000000	3.353622
take_v	10.000000	4.399693
feed_v	9.000000	3.013807
come_v	9.000000	4.205209
include_v	8.000000	3.167440
stay_v	8.000000	4.776290
kill_v	8.000000	5.667242
cover_v	8.000000	4.648049

Source: #LancsBox – Tool: *Words.*

Table 30 – Most frequent adjectives in *Animals* subset

▼ Corpus Corpus 4 - Animals ▼ Frequency ▼ Dispersion ▼ Lemma

Lemma	Frequency: 01 - Freq	Dispersion: 01_CV
many_adj	34.000000	2.054967
other_adj	32.000000	2.245188
large_adj	31.000000	2.080931
more_adj	26.000000	2.680737
small_adj	22.000000	2.530242
different_adj	21.000000	2.600437
long_adj	20.000000	2.703013
most_adj	16.000000	3.581773
young_adj	16.000000	3.217270
wild_adj	13.000000	4.102244
strong_adj	11.000000	3.138454
big_adj	11.000000	3.047289
common_adj	10.000000	4.392183
same_adj	9.000000	3.660676
sharp_adj	9.000000	4.007142
several_adj	8.000000	3.683311
only_adj	8.000000	3.615619
hard_adj	8.000000	3.872091
nocturnal_adj	7.000000	7.874008
able_adj	7.000000	3.728278
cold-blooded_adj	7.000000	4.288046
special_adj	6.000000	3.604068
soft_adj	6.000000	7.874008
african_adj	6.000000	6.564958
awake_adj	6.000000	7.874008
dry_adj	6.000000	4.369096
domestic_adj	6.000000	5.200150
pet_adj	6.000000	6.288279
warm-blooded_adj	6.000000	5.565043
poisonous_adj	5.000000	5.661440
such_adj	5.000000	5.080875
active_adj	5.000000	5.512218

Source: #LancsBox - Tool: *Words.*

Table 31 – Most frequent adverbs in *Animals* subset

▼ Corpus Corpus 4 - Animals ▼ Frequency ▼ Dispersion ▼ Lemma

Lemma	Frequency: 01 - Freq	Dispersion: 01_CV
also_adv	28.000000	2.320397
how_adv	22.000000	2.743156
not_adv	20.000000	3.075734
when_adv	20.000000	3.501384
about_adv	18.000000	2.899316
often_adv	16.000000	2.598150
usually_adv	13.000000	3.084852
sometimes_adv	13.000000	2.647392
very_adv	12.000000	2.762805
long_adv	11.000000	3.598514
most_adv	11.000000	3.962120
even_adv	11.000000	3.100704
out_adv	11.000000	3.379487
all_adv	9.000000	3.981804
together_adv	9.000000	3.689670
up_adv	9.000000	3.395959
where_adv	8.000000	2.964043
almost_adv	7.000000	3.476795
however_adv	7.000000	4.211110
why_adv	6.000000	4.080658
instead_adv	6.000000	4.281289
only_adv	6.000000	4.489758
closely_adv	6.000000	3.765027
so_adv	5.000000	4.444912
too_adv	5.000000	6.215272
everywhere_adv	5.000000	4.260263
as_adv	5.000000	4.651039
away_adv	4.000000	5.022846
just_adv	4.000000	4.517656
then_adv	4.000000	5.301937
there_adv	4.000000	5.221658
down_adv	4.000000	4.582279

Source: #LancsBox - Tool: *Words.*

4.2.4.2 N-grams tool

The approach to the contents of Table 32 is the same as to the contents of Table 19 with the teachers selecting the n-grams at their discretion. She considered the *usefulness* and *practicality* of the language patterns not only for the implementation of tasks in the lessons, but also their impact on learners' future fluency when using the language.

Table 32 – Most frequent 3- and 4-gram clusters in the *Animals* subset

▼ Corpus Corpus 4 - Animals ▼ Frequency ▼ Dispersion ▼ Type ▼ Corpus Corpus 4 - Animals ▼ Frequency ▼ Dispersion ▼ Type

Type	Frequency: 01 - Freq	Dispersion: 01	Type	Frequency: 01 - Freq	Dispersion: 01
species or types	18.000000	3.562577	species or types of	15.000000	3.695094
answer the questions	16.000000	2.242937	directions read the text	10.000000	2.956084
or types of	15.000000	3.679384	there are more than	10.000000	3.899956
there are about	14.000000	3.609469	the animals that are	9.000000	2.878753
animals that are	11.000000	2.708282	species or kinds of	8.000000	3.687922
read the text	11.000000	2.752182	and answer the questions	7.000000	3.457189
there are more	10.000000	3.896570	the text and answer	6.000000	3.579115
are more than	10.000000	3.896570	read the text and	6.000000	3.579115
the animals that	10.000000	2.821383	text and answer the	6.000000	3.579115
directions read the	10.000000	2.946004	text answer the questions	5.000000	3.544917
that live in	9.000000	4.347907	read the text answer	5.000000	3.544917
species or kinds	9.000000	3.456406	is an animal that	5.000000	3.962015
during the day	8.000000	6.639977	the text answer the	5.000000	3.544917
or kinds of	8.000000	3.684578	everywhere in the world	5.000000	4.266654
are animals that	8.000000	3.163193	circle the animals that	5.000000	3.506314
live in the	8.000000	3.373748	the words in the	4.000000	4.680695
live on land	8.000000	5.488024	vertebrates vertebrates have backbones	4.000000	7.874008
belong to the	7.000000	5.105445	lungs live on land	4.000000	7.874008
known for their	7.000000	3.618910	in the box to	4.000000	4.680695
are insects that	7.000000	6.340722	that live in the	4.000000	4.903465
animals that have	7.000000	3.514612	use the words in	4.000000	4.680695
in the world	7.000000	3.358339	this means that they	4.000000	5.580129
and answer the	7.000000	3.453719	today there are about	4.000000	5.716101
means that they	6.000000	3.974373	are animals that have	4.000000	4.450800
an animal that	6.000000	3.489321	are known for their	4.000000	4.169869
are the largest	6.000000	5.055516	they are related to	4.000000	4.316469
have a backbone	6.000000	4.159412	words in the box	4.000000	4.680695
the text and	6.000000	3.574719	of their lives in	4.000000	4.748158
text and answer	6.000000	3.574719	earth long before humans	3.000000	6.004001
group of animals	6.000000	5.055647	earth for more than	3.000000	6.201262
what is a	6.000000	5.243389	100,000 species or types	3.000000	5.754360
part of the	6.000000	4.689009	are constantly discovering new	3.000000	6.004001

Source: #LancsBox – Tool: *Ngram.*

4.2.5 Choice of vocabulary in *Animals* (T2)

The processes of language analysis with the subset *Animals* were the same as those used with the subset *Neighborhood*. In addition to the word classes, the teacher also identified the frequency of word occurrences. One of the words selected was *animals* to be used compared with *animal*, the former in 11th place and the second in the 24th place in the general frequency list of the corpus *Animals* (Table 33).

4.2.5.1 *KWIC* and *Words* tools

The concordance lines with *animals* (Figure 39) were abounding which made the task of selecting different ones for the individual learners much easier. The concordance lines also mentioned different types of animals, parts of their bodies, the baby animals as well as the different ways they move, areas focused in the classwork and homework.

Figure 39 – Noun: *animals*

Search animals **Occurrences** 133 (137.74) Texts 41/63 ▼ **Corpus** Corpus 4 - Animals ▼ **Contex**

Ind...	File	Left	Node	Right
4	LifeSci_1stG_B:	Look at the pictures. Match the baby	animals	to the parents. Draw lines to connect
5	LifeSci_1stG_T(	Read the text. Answer the questions. Some	animals	have shells. Shells help protect them. You
6	LifeSci_3rdG_A	for One and One for All Many	animals	live in groups. There are many reasons
7	LifeSci_3rdG_A	help each other care for young. Some	animals	will watch for predators while others eat.
8	LifeSci_3rdG_A	for predators while others eat. Some grown-up	animals	live with their young to keep them
9	LifeSci_3rdG_A	keep them safe and teach them. Some	animals	hunt together. This helps the whole group
10	LifeSci_3rdG_A	enough food. Living in a group helps	animals.	1. Why do some animals live in
11	LifeSci_3rdG_A	group helps animals. 1. Why do some	animals	live in groups? to help each other,
12	LifeSci_3rdG_A	young and to hunt together 2. All	animals	in a group help each other Some
13	LifeSci_3rdG_A	in a group help each other Some	animals	stand in a circle. Stronger animals stand
14	LifeSci_3rdG_A	Some animals stand in a circle. Stronger	animals	stand on the outside of the circle,
15	LifeSci_3rdG_A	the outside of the circle, and weaker	animals	stand inside the circle. 1. What animals
16	LifeSci_3rdG_A	animals stand inside the circle. 1. What	animals	stand on the outside of the circle?
17	LifeSci_3rdG_A	on the outside of the circle? baby	animals	2. Why might the weaker animals stand
18	LifeSci_3rdG_A	baby animals 2. Why might the weaker	animals	stand inside the circle? for protection gaggle
19	LifeSci_3rdG_A	lions herd of deer 1. Groups of	animals	sometimes have the same name. 2. What
20	LifeSci_3rdG_A	more than the other. Another job of	animals	is to help each other stay clean.
21	LifeSci_3rdG_A	males. 1. What is one job male	animals	often have? protecting the group 2. What
22	LifeSci_4thG_A	the text, and answer the questions. Nocturnal	Animals	Some animals sleep during the day and
23	LifeSci_4thG_And	answer the questions. Nocturnal Animals Some	animals	sleep during the day and are awake
24	LifeSci_4thG_A	day and are awake at night. These	animals	are called nocturnal. Animals that sleep at
25	LifeSci_4thG_A	at night. These animals are called nocturnal.	Animals	that sleep at night and are awake
26	LifeSci_4thG_A	during the day are called diurnal. Many	animals	are nocturnal. They eat and are active
27	LifeSci_4thG_A	nocturnal animal hunt, eat, are active Nocturnal	animals	have eyes that are extra sensitive to
28	LifeSci_4thG_A	for food and fly Cats are nocturnal	animals	because they prefer to be active at
29	LifeSci_4thG_A	Vertebrates and Invertebrates Some	animals	have backbones. They are called vertebrates. Birds,
30	LifeSci_4thG_A	protected by skeletal framework called a skull.	Animals	without backbones are invertebrates. Spiders, insects, and
31	LifeSci_4thG_A	They lay eggs and they have scales.	Animals	with backbones tend to be faster and
32	LifeSci_4thG_A	tend to be faster and stronger than	animals	without backbones. One of the functions of

Source: #LancsBox – Tool: *KWIC*.

Table 33 – Noun: *animals* – distribution in texts and relative frequency per 10k

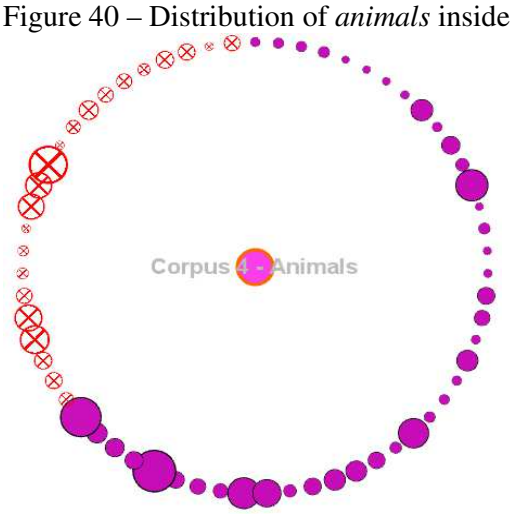
Corpus 4 - Animals: animals

File	Tokens	Frequency	▼ Relative fre
Science_K_Animals_25.txt	40	7	1750.0
Science_K_TameandWildAnima...	48	7	1458.3333
Science_K_HerbivoresandCarni...	49	5	1020.4082
Science_K_Omnivores_16.txt	56	5	892.8572
Science_PreK_YoungAnimals_...	25	2	800.0
Science_PreK_Fish_04.txt	27	2	740.7407
Science_PreK-Reptiles_05.txt	28	2	714.28577
Science_PreK_Amphibians_06.txt	32	2	625.0
Science_3rdG_VertebratesvsInv...	205	10	487.80487
Science_3rdG_Invertebrates_16...	42	2	476.1905
LifeSci_K_AdultAnimalsandBabi...	146	6	410.9589
LifeSci_1stG_BabyAnimals_W9...	76	3	394.73682
LifeSci_3rdG_Animals-LivinginG...	443	16	361.17383
Science_PreK_Mammals_02.txt	28	1	357.14288
Sci - 3rdG - Insects - Britannica....	57	2	350.8772
LifeSci_1stG_BabyAnimalsV_W...	29	1	344.82758
Science_PreK_Birds_03.txt	31	1	322.58063
Science 3rdG_TypesofAnimals_...	133	4	300.7519
LifeSci_K_AdultAnimals_W10D...	107	3	280.3738

Source: #LancsBox – Tool: *Words.*

In future implementations in the classroom, a further benefit for learners is that the teacher is able to identify the file(s) where she can find a larger number of the *KWIC of interest* using the Table 33 above. As an example, she can work with the word *animals* using the files where the number of tokens is higher, either LifeSci_3rG_Animals-LivingG with 443 tokens, or Science_3rdG_VertebratesvsIn with 205 tokens. Those files would offer recurrence of exposure and repetition - *input flooding* - of the KWIC affording higher noticing of the node and its neighbors.

Figure 40 – Distribution of *animals* inside the corpus and location in the files



Source: #LancsBox – Tool: *Words.*

The image above shows an overpopulated corpus where the names of the files do not appear immediately; one has to right click on the dot of interest and the name will be visible. This is another way to identify the most interesting files which can be used by learners themselves in a future class. The bigger and darker the dots, the higher the number of the target language in those files.

Figure 41 – Noun: *animal*

Search animal **Occurrences** 45 (46.60) Texts 29/63 ▼ **Corpus** Corpus 4 - Animals

Index	File	Left	Node	
1	LifeSci_1stG_	the chart. Answer the questions. Type of	Animal	How Long They Stay with Their Mothers
2	LifeSci_3rdG_	attack the lion. What's Your Job? Each	animal	in a group has a job. These
3	LifeSci_4thG_	and wolves are nocturnal. 1. A nocturnal	animal	sleeps during the day 2. During the
4	LifeSci_4thG_	the day 2. During the night, nocturnal	animal	hunt, eat, are active Nocturnal animals have
5	LifeSci_4thG_	help protect them. invertebrate vertebrate 1. Which	animal	is an invertebrate? ladybug 2. All vertebrates
6	LifeSci_4thG_	an animal's backbone flexible. They allow an	animal	to move freely. If the backbone were
7	LifeSci_4thG_	the backbone were one solid bone, the	animal	would be stiff. It couldn't walk or
8	LifeSci_K_Adu		Animal	Bodies All animals have body parts they
9	LifeSci_K_Adu	fly? wing 2. What part of an	animal	helps it hear? ears eye tail mouth
10	LifeSci_K_Adu	you to school 2. What does an	animal	parent do for its babies? feed them
11	LifeSci_K_Adu	a cub 2. How is the baby	animal	different from its parent? It has less
12	LifeSci_K_Adu	has less hair. 3. Draw a baby	animal	and its parent. Baby Animals Babies don't
13	LifeSci_K_Adu	different color. But babies are the same	animal.	1. Baby animals don't look exactly like
14	LifeSci_K_Adu	Sometimes baby animals are a different color	animal	than their parents Baby animals have different
15	Sci - 3rdG - Er	species is any type of plant or	animal	that is in danger of disappearing forever.
16	Sci - 3rdG - Er	a species, or type, of plant or	animal	dies out completely, it becomes extinct.
17	Sci - 3rdG - Fe	cheetah The cheetah is the fastest land	animal	on Earth. This spotted member of the
18	Sci - 3rdG - Ma	ass are used to identify the same	animal.	However, the term donkey is used for
19	Sci - 3rdG - Ro	many countries they are the most common	animal.	porcupine Porcupines are rodents with sharp spines
20	Sci - 4thG - Fis	the sharks. Fish are a kind of	animal	that lives in water. Fish have lived
21	Sci - 4thG - Ma	MAMMALS A mammal is an	animal	that breathes air, has a backbone, and
22	Sci - 4thG - Ma	most highly developed. A mammal is an	animal	that breathes air, has a backbone, and
23	Sci - 4thG - Mo	shell. A mollusk is a kind of	animal	with a soft body. Most mollusks have
24	Sci - 4thG - Re	Reptiles A reptile is an air-breathing	animal	that has scales instead of hair or
25	Sci - 4thG - Re	coiling around it so tightly that the	animal	cannot breathe. chameleon chameleon The lizards called
26	Sci - 4thG- Oth	hare. They all belong to the same	animal	family,... skunk skunk Skunks are black and
27	Sci - 5thG - Ins	Like other parasites, fleas depend on the	animal	they live on for food. Fleas bite
28	Science_1stG	enemies. The colors or patterns that an	animal	has can help it hide from dangerous
29	Science_2ndG	phenomenon unique to the shark. No other	animal	in the world has teeth quite like
30	Science_2ndG	endangered species is a plant or an	animal	that is in danger of becoming extinct,
31	Science_3rdG		Animal	Groups Animals are divided into groups that

Source: #LancsBox – Tool*: KWIC.*

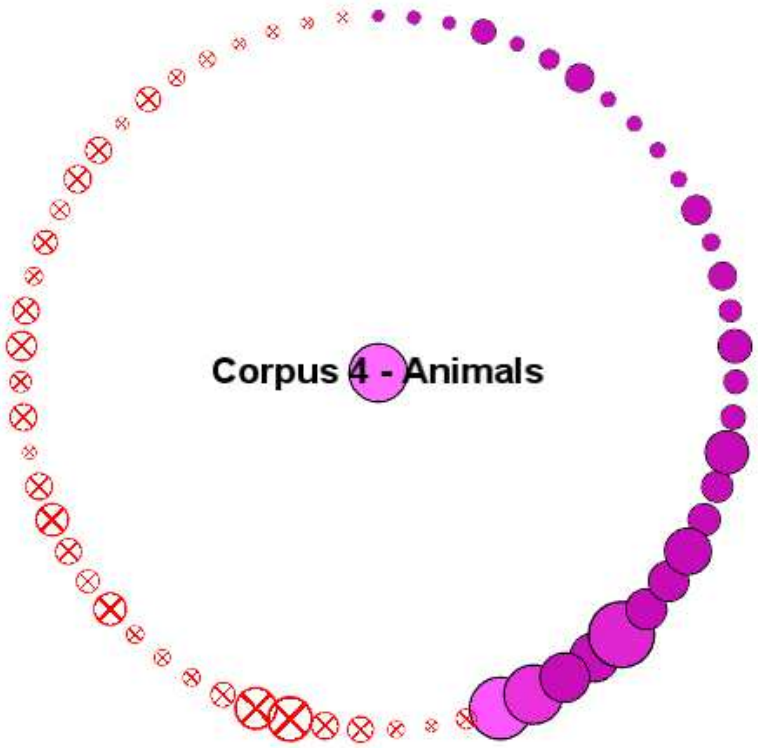
Table 34 – Noun: *animal* – distribution in texts and relative frequency per 10k

Corpus 4 - Animals: animal

File	Tokens	Frequency	▼ Relative freq
Science_PreK_YoungAnimals_...	25	2	800.0
Sci - 3rdG - Endangered specie...	33	2	606.0606
Science_PreK_Birds_03.txt	31	1	322.58063
LifeSci_K_AdultAnimals_W10D...	107	3	280.3738
Science_3rdG_Vertebrates_15.txt	36	1	277.77777
Science_K_Pets_27.txt	72	2	277.77777
LifeSci_K_AdultAnimalsandBabi...	146	4	273.9726
Science_K_Animals_25.txt	40	1	250.0
Science_2ndG_EndangeredSpe...	41	1	243.90244
Science_3rdG_Invertebrates_16...	42	1	238.09525
Science_3rdG_Predators_18.txt	46	1	217.39131
Science_3rdG_AnimalGroups_2...	155	3	193.54839
Science_K_Omnivores_16.txt	56	1	178.57144
Sci - 4thG - Mammals - Britannic...	141	2	141.84396
LifeSci_1stG_Animals-Parentin...	91	1	109.890114
Science_3rdG_VertebratesvsInv...	205	2	97.560974
Science_2ndG_Animals-Sharks...	104	1	96.15385
Sci - 4thG - Fish - Britannica.doc...	109	1	91.74312
LifeSci_4thG_Animals-Vertebrat...	340	3	88.2353

Source: #LancsBox – Tool: *Words.*

Figure 42 – Distribution of *animal* inside the corpus and location in the files



Source: #LancsBox – Tool: *Words.*

Once more the concordance lines yielded by the concordancer were copied and enlarged to make it easier for the youngsters to observe the KWICs and the surrounding context (Figure 43). Clara used different approaches to call learners' attention to segments of the sentences that could be useful when they produced their own sentences to describe the animals in their homework and posttests. Some examples are *jump very well, jumping insects, animals that have strong legs to run, have feathers, … that cannot fly* and *cannot fly at all*. Learners worked with them while looking at images in the pptx, handling concordance lines like the ones below and playing *bingo* (Appendix I).

Figure 43 – Samples of the assorted concordance lines selected - *Animals*

Frogs are small animals that can jump very well.
Many people have birds as pets.
Crickets are jumping insects.
Mosquitos are insects that are found almost everywhere in the world.
The tiger is the largest of the cats.
Birds are animals that have feathers. They lay eggs.

Zebras are animals that have strong legs to run.
Penguins are birds that cannot fly.
Some birds cannot fly at all.

Source: #LancsBox – Tool: *KWIC.*

4.2.5.2 *Text* tool

Figure 39 (subsection 4.2.5.1 above) shows the 31 concordance lines with the KWIC *animals* in red. The names of those files on the left of the image, next to the concordance lines, indicate where the KWIC is situated. Once aware of that, the teacher can open the whole text by clicking on the node *animals* in the concordance line chosen and a pop-up would open with it.[58] In just one text the learners would be exposed to an *input flooding* (Sharwood Smith, 1993) of 12 tokens of the target word which could afford more noticing and effective retention of the target language (Figure 44).

Figure 44 – Examples of node "animals" in the concordance lines

Many animals live in groups. There are many reasons for this. They live in groups for safety. They live in groups to help each other care for young. Some animals will watch for predators while others eat. Some grown-up animals live with their young to keep them safe and teach them. Some animals hunt together. This helps the whole group get enough food. Living in a group helps animals.
1. Why do some animals live in groups? to help each other, care for safety for young and to hunt together
2. All animals in a group help each other
Some animals stand in a circle. Stronger animals stand on the outside of the circle, and weaker animals stand inside the circle.
1. What animals stand on the outside of the circle? baby animals
2. Why might the weaker animals stand inside the circle? for protection

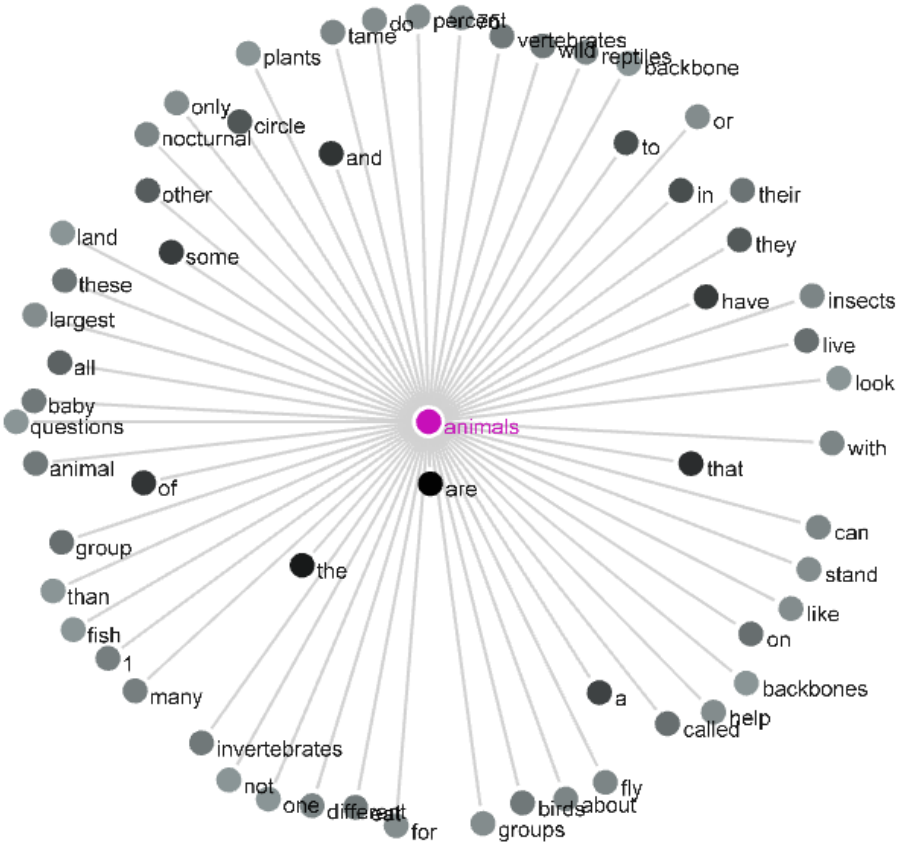
Source: #LancsBox – Tool: *Text.*

4.2.5.3 *GraphColl* tool

As mentioned in the previous section in relation to the KWIC *neighborhood* and its analysis by the *GraphColl* tool, activities with visually attractive collocates of *animals* can afford effective noticing by learners. One example is sentence building in groups, which can motivate and challenge them while bringing many benefits (Chart 2).

[58] The text was copied from the software to fit this page and pasted.

Chart 2 – Animals and its 60 collocates



Source: #LancsBox – Tool: *GraphColl.*

Table 35 – The first 21 collocates and its positions and frequency around the node

animals

Freq: 133 - Collocates: 60

Index	Status	Position	Collocate	▼ Stat	Freq (coll.)	Freq (corpus)
1	○	M	are	86.0	86	322
2	○	L	the	56.0	56	445
3	○	R	that	39.0	39	150
4	●	-	animals	38.0	38	133
5	○	L	and	34.0	34	240
6	○	L	of	33.0	33	231
7	○	R	have	30.0	30	141
8	○	L	some	29.0	29	50
9	○	R	a	26.0	26	238
10	○	R	in	21.0	21	163
11	○	R	to	21.0	21	168
12	○	L	circle	18.0	18	16
13	○	R	they	17.0	17	166
14	○	L	other	16.0	16	32
15	○	L	all	14.0	14	55
16	○	R	called	11.0	11	46
17	○	L	group	11.0	11	26
18	○	R	live	11.0	11	65
19	○	R	on	11.0	11	57
20	○	R	their	10.0	10	110
21	○	L	these	10.0	10	22

Source: #LancsBox – Tool*: GraphColl*.

Using the circle in Chart 2, one suggestion is that the teacher chooses *are* or *that,* for example, and right-click on the dot next to the word. The concordance lines pop-up and learners can see the verb *are* or the adverb *that* in blue and their collocation in relation to the noun *animals* (Figures 45-46). Most of the collocates are on the right of the node but there are many on the left as well. As an example of an activity, the class can be divided into groups which will work with one of the quadrants of the image (Chart 2) and discover the most recurrent patterns.

Figure 45 – Pop-up showing the position of collocate *are*

KWIC: animals > are

Search animals **Occurrences** 74/133 (76.64) Texts 30/63 ▼ **Corpus** Corpus 4 - Animals ▼ **Context** 7 ▼ **Display Text**

Index	File	Left	Node	Right
3	LifeSci_1stG_E	How do baby animals compare to adult	animals?	similar 2. How are kittens different from
6	LifeSci_3rdG_A	for One and One for All Many	animals	live in groups. There are many reasons
24	LifeSci_4thG_A	day and are awake at night. These	animals	are called nocturnal. Animals that sleep at
25	LifeSci_4thG_A	at night. These animals are called nocturnal.	Animals	that sleep at night and are awake
26	LifeSci_4thG_A	during the day are called diurnal. Many	animals	are nocturnal. They eat and are active
27	LifeSci_4thG_A	nocturnal animal hunt, eat, are active Nocturnal	animals	have eyes that are extra sensitive to
28	LifeSci_4thG_A	for food and fly Cats are nocturnal	animals	because they prefer to be active at
29	LifeSci_4thG_A	Vertebrates and Invertebrates Some	animals	have backbones. They are called vertebrates. Birds,
30	LifeSci_4thG_A	protected by skeletal framework called a skull.	Animals	without backbones are invertebrates. Spiders, insects, and
36	LifeSci_K_Adul	Answer the questions. 1. What kind of	animals	are these? a lion and a cub
39	LifeSci_K_Adul	exactly like their parents 3. Sometimes baby	animals	are a different color animal than their
42	Sci - 3rdG - Ins	The insects are the largest group of	animals.	In fact, about 75 percent of all
43	Sci - 3rdG - Ins	In fact, about 75 percent of all	animals	are insects. Insects developed on Earth long
44	Sci - 3rdG - Mar	meat, milk, and wool. They are hardy	animals	that can live on coarse, thin grass.
45	Sci - 3rdG - Mar	among the most valuable of all domestic	animals.	Domestic animals are ones that have been
46	Sci - 3rdG - Mar	most valuable of all domestic animals. Domestic	animals	are ones that have been tamed for
47	Sci - 4thG - Am	fairly common... frog frog Frogs are small	animals	that can jump very well. Frogs are
49	Sci - 4thG - Am	lizards, they... toad toad Toads are small	animals	often confused with frogs. Toads, however, have
50	Sci - 4thG - Am	Amphibians and reptiles are two groups of	animals	that share certain features. They are vertebrates,

Source: #LancsBox – Tool: *GraphColl.*

Figure 46 – Pop-up showing the position of collocate *that*

Search animals **Occurrences** 39/133 (40.39) Texts 24/63 ▼ **Corpus** Corpus 4 - Animals ▼ **Context** 7

Index	File	Left	Node	Right
24	LifeSci_4thG_A	day and are awake at night. These	animals	are called nocturnal. Animals that sleep at
25	LifeSci_4thG_A	at night. These animals are called nocturnal.	Animals	that sleep at night and are awake
27	LifeSci_4thG_A	nocturnal animal hunt, eat, are active Nocturnal	animals	have eyes that are extra sensitive to
44	Sci - 3rdG - Mar	meat, milk, and wool. They are hardy	animals	that can live on coarse, thin grass.
45	Sci - 3rdG - Mar	among the most valuable of all domestic	animals.	Domestic animals are ones that have been
46	Sci - 3rdG - Mar	most valuable of all domestic animals. Domestic	animals	are ones that have been tamed for
47	Sci - 4thG - Am	fairly common... frog frog Frogs are small	animals	that can jump very well. Frogs are
48	Sci - 4thG - Am	through their skin that can kill other	animals,	including humans. The frogs live in the
50	Sci - 4thG - Am	Amphibians and reptiles are two groups of	animals	that share certain features. They are vertebrates,
51	Sci - 4thG - Ara	scorpion Scorpions are small	animals	with a curved tail that can deliver
52	Sci - 4thG - Ara	and mite Ticks and mites are tiny	animals	that are found all over the world.
53	Sci - 4thG - Arth	Arthropods are	animals	that have a hard outside covering called
56	Sci - 4thG - Arth	and bees are all insects. Crustaceans are	animals	that usually have a hard covering, or
58	Sci - 4thG - Bird	them unique since birds are the only	animals	that do have feathers. From pigeons in
59	Sci - 4thG - Bird	In fact, birds are the only living	animals	that have feathers. Birds have fascinated people
60	Sci - 4thG - Fish	FISH Fish are	animals	that live in the fresh and salt
69	Sci - 4thG- Othe	The biggest bears are the world's largest	animals	that live on land and eat meat.
76	Science 3rdG_	MAMMALS Mammals include humans and all other	animals	that are warm-blooded vertebrates (vertebrates have backbones)
82	Science_2ndG_	Arachnids Arachnids are	animals	that have two main body parts, and
84	Science_3rdG_	Animal Groups	Animals	are divided into groups that share key
86	Science_3rdG_	animals on Earth are invertebrates. Circle the	animals	that are invertebrates. Eel Clam Whale Dolphin
87	Science_3rdG_	predator is an animal that kills other	animals	for food. Draw a line between each
88	Science_3rdG_	are all vertebrates. Circle all of the	animals	that are vertebrates. Bird Frog Lobster Jelly
93	Science_3rdG_	not those animals have a backbone. Vertebrates	Animals	that have a unique backbone fit into
97	Science_3rdG_	the category of invertebrates. These are the	animals	that do not have a backbone. Animals
104	Science_K_Ani	walk and run. Some animals swim. Some	animals	fly. Animals that fly have wings. Circle
105	Science_K_Ani	run. Some animals swim. Some animals fly.	Animals	that fly have wings. Circle each animal
109	Science_K_He	are called carnivores. Circle all of the	animals	that are herbivores. Point to the animals
110	Science_K_He	animals that are herbivores. Point to the	animals	that are carnivores and say their names
122	Science_K_Tar	house. These animals are tame. Circle the	animals	that are wild. Point to the animals
123	Science_K_Tar	animals that are wild. Point to the	animals	that are tame and can be kept
124	Science_PreK-	Reptiles Reptiles are land	animals	that have dry skin covered in scales.
125	Science_PreK-	scales. They also lay eggs. Circle the	animals	that are reptiles. alligator snake squirrel turtle
126	Science_PreK_	Amphibians Amphibians are	animals	that live in water when they are
127	Science_PreK_	older, they live on land. Circle the	animals	that are amphibians. raccoon frog penguin salamander
128	Science_PreK_	Birds Birds are	animals	that have feathers. They also lay eggs.
129	Science_PreK_	Fish Fish are	animals	that live under water. They have fins

Source: #LancsBox – Tool: *GraphColl.*

4.2.5.4 *N-grams* tool

The last tool used was *Ngrams* which yielded the most frequent 3- and 4-gram clusters of language (Tables 36-37). Once more, there are clusters with directions to learners and also phrase fragments which were used at the discretion of the teacher (Appendix O).

Table 36 – 3-gram clusters

▼ Corpus Corpus 4 - Animals ▼ Frequency ▼ Dispersion ▼ Type

Type	▼ Frequency: 01 - Freq	Dispersion: 01_CV
species or types	18.000000	3.562577
answer the questions	16.000000	2.242937
or types of	15.000000	3.679384
there are about	14.000000	3.609469
animals that are	11.000000	2.708282
read the text	11.000000	2.752182
there are more	10.000000	3.896570
are more than	10.000000	3.896570
the animals that	10.000000	2.821383
directions read the	10.000000	2.946004
that live in	9.000000	4.347907
species or kinds	9.000000	3.456406
during the day	8.000000	6.639977
or kinds of	8.000000	3.684578
are animals that	8.000000	3.163193
live in the	8.000000	3.373748
live on land	8.000000	5.488024
belong to the	7.000000	5.105445
known for their	7.000000	3.618910
are insects that	7.000000	6.340722
animals that have	7.000000	3.514612

Source: #LancsBox – Tool: Ngrams.

Table 37 – 4- gram clusters

▼ Corpus Corpus 4 - Animals ▼ Frequency ▼ Dispersion ▼ Type

Type	▼ Frequency: 01 - Freq	Dispersion: 01_CV
species or types of	15.000000	3.695094
directions read the text	10.000000	2.956084
there are more than	10.000000	3.899956
the animals that are	9.000000	2.878753
species or kinds of	8.000000	3.687922
and answer the questions	7.000000	3.457189
the text and answer	6.000000	3.579115
read the text and	6.000000	3.579115
text and answer the	6.000000	3.579115
text answer the questions	5.000000	3.544917
read the text answer	5.000000	3.544917
is an animal that	5.000000	3.962015
the text answer the	5.000000	3.544917
everywhere in the world	5.000000	4.266654
circle the animals that	5.000000	3.506314
the words in the	4.000000	4.680695
vertebrates vertebrates have backbones	4.000000	7.874008
lungs live on land	4.000000	7.874008
in the box to	4.000000	4.680695
that live in the	4.000000	4.903465
use the words in	4.000000	4.680695
this means that they	4.000000	5.580129
today there are about	4.000000	5.716101
are animals that have	4.000000	4.450800

Source: #LancsBox – Tool: *Ngrams.*

Subsection 4.2 part (a) addressed the investigation's results showing the relevance of compiling pedagogic corpora for younger learners and the language analysis of a concordancer. The basic functions of each of the concordancer tools were described to show the relevance technology can have to enhance learning a language in the classroom. All the findings can add variety and challenge to the youngsters' classes while they engage in the new material and manipulate authentic language meaningfully and playfully. It also brings the pedagogic corpora to the forefront among materials for young learners. The next subsections 4.3 and 4.4 will address Part B of the study's outcomes to illustrate the quantitative and qualitative results, respectively, of the implementation of the treatment in all groups of learners.

Part B

4.3 Quantitative results

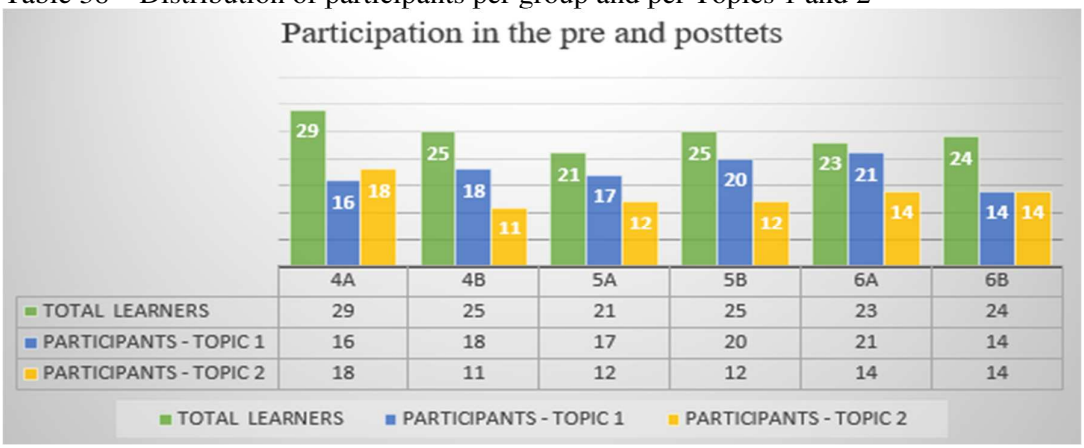
In this second part of the investigation's results, the quantitative analysis of the findings of pre-tests and posttests will be presented and analyzed. Two types of measurement tests were used to compare the findings to shed further light on question (iii) and address question (iv):

(iii) Can activities implemented with a Data-driven learning (DDL) approach expand learners' topicalized vocabulary to boost their progress in English?; and

(iv) Are the results significantly different from one grade to the others when the same tasks are worked with in the classrooms?

For statistical purposes, it was decided that students who had missed one or more classes where the investigation was carried out would have their names excluded from the final count. Thus, the final number of participants in pre and posttests in both Topic 1 and Topic 2 was 106, 72% of the total 147 students, in all the six classes (Table 38).

Table 38 – Distribution of participants per group and per Topics 1 and 2



	4A	4B	5A	5B	6A	6B
TOTAL LEARNERS	29	25	21	25	23	24
PARTICIPANTS - TOPIC 1	16	18	17	20	21	14
PARTICIPANTS - TOPIC 2	18	11	12	12	14	14

TOTAL LEARNERS PARTICIPANTS - TOPIC 1 PARTICIPANTS - TOPIC 2

Source: the researcher.

4.3.1 Analysis tools and results

The first one, the *t*-test,[59] a parametric test, was used to determine whether the treatment, work with the target language (KWICs) in concordance lines, had had a positive effect on the students' language production and establish if and how much the groups differed from one another. Since the data had unequal variances, i.e., did not entirely fit the assumptions made for the *t*-test, a second test, the Wilcoxon Rank Sum, a non-parametric test, was used to complement the analysis of the findings.

The final data was initially processed by the SPSS software[60] and a complementary analysis was carried out by the Wilcoxon test. The findings of the statistical analyses offer insights into the work carried out in the classrooms while addressing research question (iv): Are the results significantly different from one year to the others when the same tasks are worked with in the classrooms?

The percentage of learners (Table 39) in Group 4A who used lexis from the concordance lines in the posttest increased 50% from pre-test, rising from 21% to 42%. However, learners' use of target language in groups 4B and 5A had a reduction, from 29% to 25% in 4A and from 50% to 30% in 5A. Group 5A had a similar number of students on both occasions but it did not affect the results as the use of target language in the posttest dropped noticeably showing the treatment in the classroom did not impact students' production. As for 6A, the number of learners who did the tests fell considerably from 23 to 16, showing an absence of a third of the students, probably causing the 50% drop in the production of target language. Yet, 6B had the same number of students in class and the same low percentage – 35% – in pre and posttests. The 6A and 6B groups have students with higher levels of knowledge of English which may explain their possible use of alternative language for the production task.

Table 39 – % of students who used the target language from concordance lines

	4A	4B	5A	5B	6A	6B
Number of students participating in classwork	24	24	20	19	23	17
Percentage of students who used target language from concordance lines	21%	29%	50%	31%	13%	35%
Number of students who took the Posttest	24	24	21	23	16	17
Number of students who used the target language from the concordance lines in the Posttest	42%	25%	30%	35%	5%	35%

Source: the researcher

4.3.2 Analyses of posttests results

[59]See: https://libguides.library.kent.edu/SPSS/PairedSamplestTest.
[60] IBM **SPSS**® Statistics is the world's leading statistical software used to solve business and research problems.

The classwork as well as the homework were preceded by a pre-test and followed by a posttest 10 days later in all grades. Learners worked with one Topic 1 from COREL-GEO: *Neighborhood*, and one from COREL_SCI: *Animals*, both selected by the teacher due to their correlation with the curriculum maps for English in the school.

The data collected from pre and posttests was analyzed by a Paired Samples *t* test[61] (Table 40) which is used to test if the means of two paired measurements, such as pre and posttest scores, are significantly different. It is a parametric test where the dependent variable is measured at two different times. The results of the pre and posttests (total of 27 points) were standardized on a scale of 0 to 100. The *t* tests pointed out that the work with the concordance lines in Topic 1 – *Neighborhood* – in the classroom was indeed effective. There is a considerable difference in the findings. Conversely, the same could not be observed in Topic 2 - *Animals* - where there was little or no difference in the results.

Table 40 – Data analyses and *t*-test results

GROUP	Pre-test	Posttest	Mean difference	p value
4T1	65,61	84,75	19,15	p<0,05
4T2	66,77	68,13	1,46	p>0,05
5T1	69,97	87,29	17,32	p<0,05
5T2	75,23	80,32	5,09	p<0,05
6T1	75,13	91,22	16,08	p<0,05
6T2	81,01	84,79	2,78	p>0,05

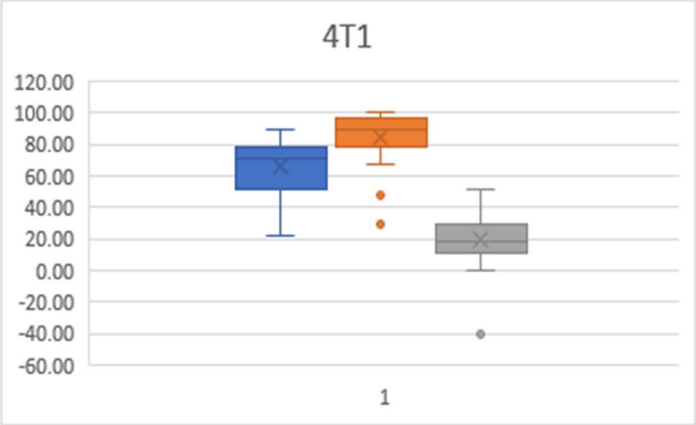
Key: T1 – Topic 1 T2 – Topic 2 4 – 4th grade 5- 5th grade 6 – 6th grade

Source: SPSS Statistics

In view of the results, boxplots were created so the data could illustrate the results of the tests further. Charts 3-4 below show the results and their distribution among the students in different grades:

[61]See: https://libguides.library.kent.edu/SPSS/PairedSamplestTest.

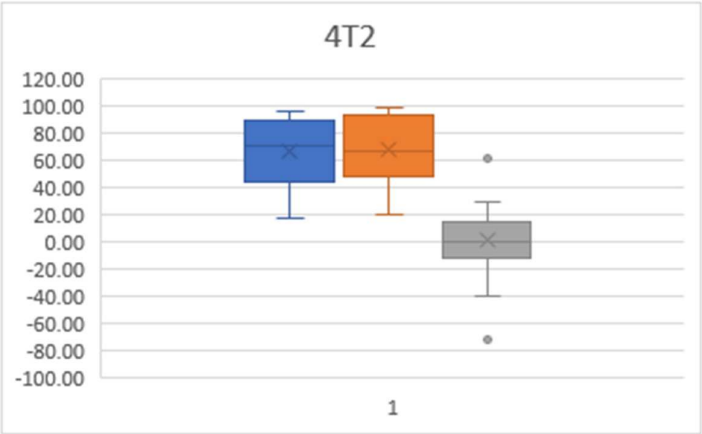
Chart 3 – 4thG A-B results – Topic 1 (T1)



Source: SPSS for Windows.

The boxplot Chart 3 illustrates the treatment effect in groups 4A and 4B in Topic 1. The blue box shows the results of pre-tests and the orange box shows the post-tests. The blue box is wider and situated at a lower position in the chart while the orange box is closer to 100 and shows less spread of participants' scores concentrated at the top. This indicates efficacy in the Topic 1 treatment and a significant result.

Chart 4 – 4thG A-B results – Topic 2 (T2)



Source: SPSS for *Windows*.

However, in Topic 2 (Chart 4), the blue and orange boxes are symmetrical although we can identify a wider upper quartile in the orange box indicating a wider spread in the values of the data and the lower quartile more condensed data. The gray box in both charts shows the difference between pre and posttests are similar. In 4T1 (4th Grade – Topic 1) there are two orange dots called outliers who could not reach the average result of the group. In 4T1 there is

a wider spread between results of the tests whereas this difference is insignificant in 4T2 (4th Grade – Topic 2).

Overall, reading the charts' findings carefully, the conclusion is that Topic 1 results are very positive unlike the results of Topic 2. The positive scores between pre- and posttests in Topic 1 had a 20% increase as a result of the significant reduction in the spread, i.e., the students had higher grades and the results were more evenly distributed across the groups. Conversely, the same scores cannot be seen in Topic 2 as the variation between the pre and post-tests ranged from 1 to 5%. Another pattern identified in Topic 1 was that pre-tests results were similar while in Topic 2 the results had a random pattern.

One explanation for groups 5G and 6G final results (Tables 41-42) is that, on average, they already had higher scores in Topic 2 pre-test than in Topic 1 pre-test. Since the top average is 85%, and they had already achieved that, it would be more difficult for the students to go beyond that result. The assumption pointed out for a confirmation of the results and further measurement was carried out. This test would check only the results of target language production (KWICs, 3- and 4-grams) Task 5 in all groups in all tests. In Topic 1 (Table 41), the 2 groups 5A and 5B did not have a significant result which had already been pointed out by the *t* test.

Table 41 – Level of significance between results of Pre and Posttests – T1

Wilcoxon Test			
Topic 1 – Pre Test – 4A	Topic 1 – Post Test – 4A	p<0,001	Sig
Topic 1 – Pre Test – 4B	Topic 1 – Post Test – 4B	p<0,001	Sig
Topic 1 – Pre Test – 5A	Topic 1 – Post Test – 5A	p=631	Not Sig
Topic 1 – Pre Test – 5B	Topic 1 – Post Test – 5B	p=161	Not Sig
Topic 1 – Pre Test – 6A	Topic 1 – Post Test – 6A	p<0,001	Sig
Topic 1 – Pre Test – 6B	Topic 1 – Post Test – 6B	p=0,008	Sig

Source: Wilcoxon Test.[2]

The results yielded insights into a possible interference of an external variable: students had more knowledge of Topic 2 (Animals) contents before the pre-test or students could relate to the contents more easily. In the analysis of Topic 2 (Table 42), however, only group 4B had a significant result:

Table 42 – Level of significance between results of Pre- and Posttests – T2

Wilcoxon Test			
Topic 2 – Pre Test – 4A	Topic 2 – Post Test – 4A	p=0,837	Not Sig
Topic 2 – Pre Test – 4B	Topic 2 – Post Test – 4B	p=0,12	Sig
Topic 2 – Pre Test – 5A	Topic 2 – Post Test – 5A	p=0,547	Not Sig
Topic 2 – Pre Test – 5B	Topic 2 – Post Test – 5B	p=0,234	Not Sig
Topic 2 – Pre Test – 6A	Topic 2 – Post Test – 6A	p=0,705	Not Sig
Topic 2 – Pre Test – 6B	Topic 2 – Post Test – 6B	p=1	Not Sig

Source: Wilcoxon Test.

In Topic 1 (Table 43) the students had a significant increase in their scores from pretest to posttest, except the 5A and 5B groups which had 6% and 10% respectively. In Topic 2 (Table 44) results show little increase in their language development, except for 4B which had a significant rise of 48% (Chart 5). Additionally, also already identified, group 6A had a decrease in the production of target language while 6B had no changes.

Table 43 – Differences between pre and posttests results – *Production Task* – Topic 1

	Topic 1 Pre Test – 4A	Topic 1 Post Test – 4A	Difference	Improvement in %
Mean	**0.18**	**2.76**	**2.59**	**86%**
	Topic 1 Pre Test – 4B	Topic 1 Post Test – 4B	Difference	Improvement in %
Mean	**0**	**2.47**	**2.47**	**82%**
	Topic 1 Pre Test – 5A	Topic 1 Post Test – 5A	Difference	Improvement in %
Mean	**2.06**	**2.22**	**0.17**	**6%**
	Topic 1 Pre Test – 5B	Topic 1 Post Test – 5B	Difference	Improvement in %
Mean	**2.5**	**2.8**	**0.30**	**10%**
	Topic 1 Pre Test – 6A	Topic 1 Post Test – 6A	Difference	Improvement in %
Mean	**0.41**	**2.55**	**2.14**	**71%**
	Topic 1 Pre Test – 6B	Topic 1 Post Test – 6B	Difference	Improvement in %
Mean	**1.36**	**2.93**	**1.57**	**52%**

Source: Wilcoxon Test.

When the results of the production task of the pre and posttests are compared in isolation (Tables 43-44), it is possible to observe significant results with increased scores in Table 43: group 4A, which rose from 0,18 to 2,76, and group 4B, from near 0 to 3. However, in Table 45, the mean of the pre and posttest in group 6B is the same 2,39, possibly because the average score of the pre-test is 2 (out of 3) which makes it more difficult to go beyond that in the pre-

test. This is the case of Topic 2, where the pre-test mean is 2,2 (out of 3) for all groups, which makes it difficult to obtain a significant rise in the posttest. The exception here is group 4B, where the mean rose from 1,27 to 2,73.

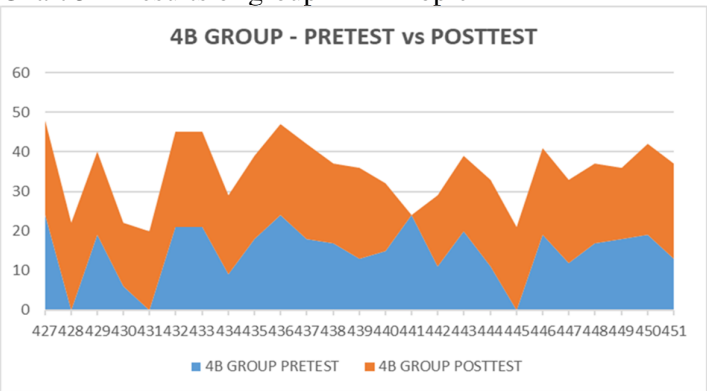
Table 44 – Differences between Pre and Posttests results – *Production Task* – Topic 2

	Topic 2 Pre Test – 4A	Topic 2 Post Test – 4A	Difference	Improvement in %
Mean	**2.42**	**2.45**	**0.03**	**1%**
	Topic 2 Pre Test – 4B	Topic 2 Post Test – 4B	Difference	Improvement in %
Mean	**1.27**	**2.73**	**1.45**	**48%**
	Topic 2 Pre Test – 5A	Topic 2 Post Test – 5A	Difference	Improvement in %
Mean	**2.38**	**2.62**	**0.23**	**8%**
	Topic 2 Pre Test – 5B	Topic 2 Post Test – 5B	Difference	Improvement in %
Mean	**1.92**	**2.42**	**0.50**	**17%**
	Topic 2 Pre Test – 6A	Topic 2 Post Test – 6A	Difference	Improvement in %
Mean	**2.67**	**2.60**	**-0.07**	**-2%**
	Topic 2 Pre Test – 6B	Topic 2 Post Test – 6B	Difference	Improvement in %
Mean	**2.39**	**2.39**	**0.00**	**0%**

Source: Wilcoxon Test.

Chart 5 below shows the difference in scores between pre- and posttests after the treatment among learners of group 4B:

Chart 5 – Results of group 4B in Topic 2



Source: the researcher.

The findings suggest that when students already have very high grades in the pre-test, they either maintain the score or show an insignificant improvement in the posttest. Further analysis of results points to the following:

1) The number of students present during the investigation was different in most groups, except groups 6B, which had the same number, and 4A that had a small increase. Those

numbers may have altered the results statistically. Table 45 shows, in the percentages, that, on average, fewer learners participated in Topic 2 classes:

Table 45 – Participants in both Topics 1 and 2

GROUPS	TOTAL LEARNERS	PARTICIPANTS TOPIC 1	% OF TOTAL	PARTICIPANTS TOPIC 2	% OF TOTAL
4A	29	16	55.17	18	62
4B	25	18	72	11	44
5A	21	17	80.95	12	57.1
5B	25	20	80	12	48
6A	23	21	91.3	14	60.86
6B	24	14	58.33	14	58.33
TOTAL	144	106	73.61	81	56.25

Source: the researcher.

2) In Topic 1, the average score of the participants had a significant rise between the pre and posttests in the following groups: 4A: 86%; 4B: 82% ; 6A: 71% and 6B: 52%. In Topic 2, just 4B had an increase of 48%.
3) In Topic 1, the average score of the participants had a non-significant increase between the tests in the following groups: 5A: 6% and 5B: 10%. In Topic 2, groups 5A: 8% and 5B: 17%.
4) The average score of the participants remained stable between the 2 tests in Topic 2 in the groups 4A: 1%; 6A: -2% and 6B: 0%.
5) The quantitative analysis points out that the language tasks have to be different for the different grades and the level of difficulty needs to be considered, information corroborated by O'Keeffe (2022). It is very important to consider the amount of time spent by each group working with English, i. e., years of exposure to the language, and also contents already dealt with in previous grades.
6) There should be more investigation towards the role of the topic contents. The findings indicate that the work with *Animals*, Topic 2, was not as effective as the work with *Neighborhood*, Topic 1, with all groups. Nonetheless, schools abroad have similar topics in all grades in *Fundamental I,* only changing the level of grammatical complexity.

The overall quantitative analysis indicates that working with KWICs and concordance lines in English in the 4th – 6th grade classrooms is not only possible but feasible. It does increase the average scores in English language tests and also expand their vocabulary

knowledge. It is still premature to answer positively to question (iii) Can activities implemented with a Data-driven learning (DDL) approach expand learners' topicalized vocabulary and possibly boost their progress in English? But as with any other approach to teaching, the answer is never that the result is 100% effective, as there are many variables which play a relevant role in the analysis and need to be taken into account.

Further work with vocabulary is recommended and assessment of retention through written performance should be tested. As for question (iv) Are the results significantly different from one year to the others when the same tasks are worked with in the classrooms?, the results show the younger groups had more benefits with the language treatment than the older groups. They clearly indicate that different grades had differences in the outcomes but, although to different degrees, the treatment had been effective in the six groups.

4.4 Qualitative analysis and results

This thesis section exemplifies the qualitative analysis of the students' written sentences which were produced in response to the activities proposed. The evidence of *noticing*, understanding of meaning and further awareness of language is clearly demonstrated by the samples in all six groups of students. The DDL underlying constructs such as inductive work, conscious awareness-raising of grammatical features of a sentence (3- and 4-grams) and recurrence of exposure are certainly present in those samples.

4.4.1 Evidence from Topic 1

In order to work with Topic 1 *Neighborhood*, the teacher received a list of possible target words in their concordance lines and selected the ones she deemed more appropriate to her students. Below are some examples of concordance lines the teacher enlarged and later cut into single sentences to be distributed to the students. Not all of them were full sentences as they were copied and pasted exactly as the concordancer generated them.

First, the learners took the pre-test (Appendix D) to check how much of the topic they already knew. The results demonstrated that although they knew most of the target words, they had difficulties writing sentences with those words in Task 5 (*production task*). Most learners either left the spaces blank or did not write a full sentence, and as a result, those were considered for statistical purposes. It served as learners' first encounter with the new vocabulary and for the teacher to gauge their interest in the topic. However, after the classwork and homework almost all the students produced full sentences in the posttest (Appendix D). The main aim of

the investigation was to expand vocabulary, not grammar, and use it in suitable sentences like the ones in the concordance lines which related to their lives.

For the classwork, Clara instructed the learners to pay attention to the KWICs and also connected them to the video used as an introductory activity. She asked them to pay attention to the target words and transfer the information to their own reality. Those lines (Figure 47) are simple sentences or parts of sentences which were used in other activities along 2 back-to-back lessons (Lesson plan in Appendix J).

Figure 47 – Some samples of concordance lines – *Neighborhood*

There are many different types of map. This is because we use different maps for different reasons.
If you were going on a hike in a nature park, you would need a map of the park.
If you wanted to know the way around your neighborhood, you would need a map of your neighborhood.
find the best road to your friend's house
find your way around a nature park
Cities are busy places! They have many buildings, including businesses and schools.
City block with homes and stores
Do you have friends in your neighborhood? Draw a picture of something you like that is near your home.
Tall apartment buildings where many people live
The part of a city or town that is close to your home is called your neighborhood.
The stores you use and your school are likely to be in your neighborhood.
They may have museums, libraries, and parks.
People live, work, learn, and have fun close to one another in cities
They may be able to walk to school, the post office, the library, and stores.
They may also use public transportation to get to different parts of the city.
City block with homes and stores
Do you have friends in your neighborhood? Draw a picture of something you like that is near your home.
Tall apartment buildings where many people live
The part of a city or town that is close to your home is called your neighborhood.

The stores you use and your school are likely to be in your neighborhood.
Many of the people you know probably live in your neighborhood.
There is a library nearby

Source: *Neighborhood* corpus. Lines yielded by #LancsBox 6.0.

The great surprise was the number of students' samples which included *nearby* in their sentences as this adverb is not used at this stage of a beginner's level. The adverb was not emphasized by the teacher but was reproduced appropriately in all groups of learners. As she used the same concordance lines with all learners, the sentences they produced became solid evidence of the teacher's work with the same material in all classes. Examples of *substitution* work with the original concordance line (A):

(A) There is a *library* nearby. (KWIC: *library*)

(B) There is a *park* nearby. (4th grade)

(C) There is a *restaurant* nearby. (5th grade)

(D) There is a *pet shop* nearby. (6th grade)

The sample sentence (A) above contains the KWIC in italics. The students substituted according to their own neighborhood (B, C, D), making it personalized, meaningful, and hopefully memorable. The learner had to understand the meaning of the sentence to substitute *school* for *park, restaurant, Mall* and so on. Video and images were used to convey meaning and enhance comprehension. The underlying aim of the concordance lines was to focus on the target language, but a secondary aim was the revision of the construction 'there + be'. It was an open exercise where students created their own sentences based on the examples of the lines.

The analysis of the learners' sentences strongly indicate they all received similar treatment, i. e., the same *noticing* work of the target language exposure to the subsets of corpora. The DDL underlying constructs such as inductive work, conscious awareness-raising of grammatical features of a sentence (3- and 4-grams) and recurrence of exposure (concordance lines) are certainly present in those samples.

Table 46 – Sample sentences from learners' posttests

Groups	Samples of students' production	Original concordance lines
4A	There is a park nearby	There is a **library** nearby
4B	There is a Mall nearby	
5A	There is a bakery nearby	
5B	There is a park nearby	
6A	There is a supermarket nearby	
6B	There is a restaurant nearby	
4A	Find the best road to your park	**Find the best road to the police station**
4B	Find the best road to your pet shop	
5A		
5B	Find the best road to your school	
6A		
6B	Find the best road to your family's building	
4A	Find your way around the park	**Find your way around a nature park**
4B	Find your way around your neighborhood	
5A	Find your way around the library	
5B	Find your way around a school	
6A	Find your way around a nature park	
6B	Find your way around a hospital	

Source: the researcher.

It also indicates learners understood the overall meaning of the KWICs[62] (*neighborhood*, *supermarket*, *hospital, school*, *police station*, *post office* (*mail office*), *bank, fire station, library, park, pharmacy (drugstore), bus stop, theater (cinema)* and the various n-grams like *find your way around*, before writing their sentences. In fact, the sentences are very similar due to the framed context learners had to work with (Table 46).

Our concern was not to deliver lessons with totally unknown contents as learners had spent a long period without proper lessons in the school. Samples of individual learners in all 6 classes are displayed in Appendix P. It is possible to observe that although the learners in the 6th grades were more creative and some even ventured away from the concordance lines using new vocabulary, some posttests were done carelessly, without attention.

4.4.2 Evidence from Topic 2

Topic 2 was introduced to all groups of learners a couple of weeks after Topic 1 was completed. This time learners were invited to work with *Animals,* to learn or revise their names, make descriptions with characteristics, body parts, and ways they moved. The approach was identical to Topic 1, with learners doing a pre-test first and in a second lesson receiving concordance lines (Table 47) to manipulate them during the activities. They also had supporting

[62] The transcript of a video which was used to introduce the topic is in Appendix H.

very colorful bingo cards which aimed not only at repetition of exposure to animal names but also to other related vocabulary.

Table 47 – Samples of concordance lines used in class and homework

Insects developed on Earth before humans did.
Butterflies, ants, and bees are all insects. Insects have 6 legs.
Butterflies have 4 wings.
Spiders are eight-legged creatures. They live in most parts of the world.
Ants and bees are social insects.
Crickets are jumping insects.
Mosquitos are insects that are found almost everywhere in the world.
Reptiles are land animals that have dry skin covered in scales.
Elephants are the largest living animals
Lizards, alligators, crocodiles, turtles and snakes are all reptiles. They lay eggs.
Alligators and crocodiles are powerful animals with powerful tails.
Fish are a kind of animal that lives in the water.
Fish have scales, lay eggs and have finns to help them swim.
Birds are animals that have feathers. They lay eggs.
Birds are vertebrates that fly flapping their wings.
Some birds cannot fly at all. Penguins are this kind of birds.
Many people have birds as pets.
Lions live in parts of Africa and India.
Zebras have strong legs to run.

Source: *Animals* corpus. Lines yielded by #LancsBox 6.0.

Those lines were enlarged and helped learners work with the language. They resorted to some of them to describe the animals and also adapted to other animals as well. Some of the groups of words were used by students: *strong legs to run, live in the water* and *jumping [...]*. Table 48 shows some samples of learners' production in the final posttest. More samples of learners' written production are in Appendix Q.

Table 48 – Sample sentences from learners' posttest

Groups	Samples of students' production	Original concordance lines
4A **4B** **5A** **5B** **6A** **6B**	Rabbits are **jumping animals.** Frogs and rabbits **are jump.** Kangaroos are big animals that **can jump very well.** Frogs are **jumping animals** Frogs are animals that **jump very well.** Rabbits are small animals **that can jump very well.**	Crickets are **jumping insects.**
4A **4B** **5A** **5B** **6A** **6B**	The duck **love water.** Shark are a kind of animals that **live in the water.** Lions are **wild animals that ive in the jungle.** Frog **live in the water.** Starfish are a **kind of animal that lives in water.** The fish **swims.**	Fish are **a kind of animal that lives in the water.**
4A **4B** **5A** **5B** **6A** **6B**	Elephants **cannot jump.** Penguins are birds that **cannot sing** Chickens are birds that **cannot fly.** Parrots are birds that **cannot swim.** Some birds **cannot fly.** Some insects **can jump.** The cat **can climb and can walk and run.**	Some birds **cannot fly at all.**
4A **4B** **5A** **5B** **6A** **6B**	Gazela are animals **that have strong legs to run.** Lions are animals that **have strong legs to run.** Giraffes **have strong legs to run.** Horses are animals that **have strong legs to run.** Zebras **have strong legs to run.** Tiger **has strong legs to run.**	Zebras **have strong legs to run.**
4A **4B** **5A** **5B** **6A** **6B**	Hippos **live in parts of Africa.** Lions are wils animals that **live in the jungle.** Rabbits **live in the house.** Blue parrots **live in parts of Brazil.** Giraffes **live in parts of Africa.** Giraffes **live in parts of Africa.**	Lions **live in parts of Africa and India.**

Source: the researcher.

The groups of words in red show chunks which were more frequently used by learners. Those clusters would have probably not been used by learners if the words had not been exposed in context as they were in class. As Corino (2014, p. 68) has posited, "corpus work and DDL can thus help teachers to find patterns of specialized phraseology, which are barely mentioned in the general bilingual and monolingual dictionaries used by their students."

4.4.3 Clara's concluding statements

The teacher could witness the engagement of young learners with the material, the tasks, and was positively surprised with the results. According to her, "they produced simple sentences in class without support, and this time with the KWICs." She felt rewarded with the changes and I assume she will probably keep using inductive techniques in her classrooms. She

observed that the learners carried out the tasks as engaged as usual, except for the manipulation of concordance lines which triggered their curiosity at first. They experienced some difficulty until she explained they were examples of different uses of the target language (KWICs). The awareness work using concordance lines to help learners notice (Schmidt, 1990) the KWICs showed good results and indicated new possibilities she would have for class work from then on. The innovation in materials triggered not only their curiosity but also their interest in finding out what kind of information was in the concordance lines.

In Clara's exchanges with myself, she implied that this investigation showed that younger learners should not be underestimated. Students who had been exposed to English in regular classes longer had more consistent positive results in the use of n-grams. This statement was confirmed by the quantitative results in sections 4.3. Examples such as *in front of* and next to the, studied with prepositions of place before, but not mentioned now, came up in the 4th grade learners' sentences about neighborhood:

- The pet shop *takes care of* the animals. (9 years old).
- The park *is next to the* library. (9 years old).
- The post office *is in front of* the store (9 years old).

The teacher noticed the learning effects mostly after classwork. The sentences above were produced in exercises where they either had to choose and describe one type of building (Topic 1), or select one among many others in a picture. Linguistic items such as *the bakery sells fresh bread* and *the pet shop takes care of the animals* among many others had been used by her on different occasions: when introducing the vocabulary, watching a video, or describing a picture, for example, and were reproduced in the learners' sentences (form and meaning recall). This *information recall* reminded the teacher that those contextualized chunks had been effectively retained and were easily retrieved when the situation required. Despite the apparent 10-day language benefits for learners, the vocabulary chosen for the activities work, which required awareness of meaning and personalized language production, proved to be meaningful to the young learners.

With regards to the concordance lines and n-grams, Clara recommended that teachers should use "concordance lines as full sentences, [and depending on the groups,] they should start the class working with the texts from the pedagogic corpora." She suggested that a longer stretch of text, or even two or three lines to contextualize the KWIC would be more beneficial. She ended her comments by stating she enjoyed using authentic language with her young learner groups, "I can say that before the lessons I had underestimated their potential and level

of understanding. It was reassuring to see them understand the language coming up with examples of their own on other occasions."

4.5 Final remarks

The quantitative analysis described in the previous subsection 4.3 corroborates Clara's comments and remarks. It shows some patterns in the results which were different in the 6 groups and also reveals differences when working with Topics 1 and 2. Due to the circumstances of implementing the tasks just after the reopening of the school to face-to-face classes, unexpected variables may have played a part in those findings. Two examples can illustrate that: different interests due to age differences and not challenging enough tasks to older learners. Once identified, they can raise teachers' and researchers' awareness of areas which need more careful design in future studies.

It is possible to say that research question (iii) Can activities implemented with a Data-driven learning (DDL) approach expand learners' topicalized vocabulary and possibly boost their progress in English? has been addressed by the Posttests. As the investigation demonstrated, different vocabulary other than the target vocabulary, was used mainly by older learners in the production stage. This factor in itself is already evidence of vocabulary expansion. However, more work with the target vocabulary in cross-curricular activities will afford more KWICs repetition, more exposure to language clusters and more retention to be assessed in the future. We can then consider the answer to research question (iii) positive, as students can expand their repertoire of lexis in different areas given many conditions suggested by the variables above are met. Perhaps the younger the learners, the more restricted their knowledge of English, and therefore, the more beneficial it will be if they are exposed to authentic and contextualized language from the initial grades in primary school. Not only individual KWICs but also the n-grams should certainly be included in their language work.

Question (iv) Are the results significantly different from one year to the others when the same tasks are worked with in the classrooms? was clearly answered as we described the findings in the posttests of all the groups. The sentences produced in the posttests show evidence from Topics 1 and 2 classwork and homework with the corpus-informed concordance lines. As a matter of fact, the outcomes were somehow expected as the age groups were as different as their interests and maturity.

Chapter 5 – Conclusion

The final comments on this investigation's results in its many facets can only be made after I voice my main concern over the lack of previous investigations with similar focus and characteristics to this one. Before this investigation on the use of corpora with young learners, we tried to find other similar studies in the literature to no avail.

First in Brazil, where corpus-based studies have been carried out extensively in many academia scattered around the country which were mentioned in Chapter 2. Then, the search was abroad when I chanced upon Boulton's declaration that although "[he had] been researching data-driven learning for many years now, [he knew] virtually nothing about DDL with younger learners. None in a primary school context" (2020, e-book loc., 346). Very few scholars had carried out studies with youngsters before (Sealey, 2011; Pérez-Paredes, 2020; Crosthwaite, Stell, 2020; Boulton, 2020), but they either had a different linguistic focus for the investigation or worked with older or more advanced groups of learners.

Therefore, I hope the points and issues to be raised here shed light into using corpus-based innovation in the young learners' English classroom. It can guide future projects to advance even further in the use of KWICs in concordance lines from pedagogic corpora to aid language exposure, noticing and effective retention of language patterns.

We also considered the many variables which could have compromised the research. Different interests among participants in the topics chosen, the use of similar activities for different-age learners, their being in different phases of adolescence and also their level of motivation after 2 years homebound due to the pandemic. One caveat was the use of the same activities in the three grades, which only confirmed the fact that the older students may not have been challenged enough to use the target language or the themes selected were below their expectations. Additionally, it was understood by the researcher and the teacher that results could only be compared statistically if they were obtained from identical tasks. Harley and Wang (1997) have argued that older learners are usually capable of faster initial progress in acquiring the grammatical and lexical components of an L2 due to their higher level of cognitive development and greater analytical abilities but later claimed that

> in terms of language pedagogy, it can therefore be concluded that (i) there is no single 'magic' age for L2 learning, (ii) both older and younger learners are able to achieve advanced levels of proficiency in an L2, and (iii) the general

> and specific characteristics of the learning environment are also likely to be variables of equal or greater importance (Harley; Wang, 2009, p. 170).

The statistical results (subsection 4.3) showed the younger groups had more benefits with the language treatment than the older groups, which clearly indicates that the same treatment affected the outcomes in different degrees. As O'Keeffe (2022, p. 6) has put forward, "it means differentiating tasks and data by level. Pedagogical focus at A1 and A2 needs to be on fostering language experience so that learners enhance their knowledge of what words go together, i.e., the basic slots and frames". Teachers can take a greater role in designing tasks which provide learners with a more fine-grained curation of stretches of language to afford task differentiation. In future studies, as learners move to upper grades, they can be made aware of new meanings and be challenged to create new phrasal word associations in context. Suggestions for further classwork with the KWICs and clusters with the use of #LancsBox tools are mentioned along subsections 4.2.1 – 4.2.5.3.

Another avenue for future investigation is the variable *time of day* when the different groups have the classwork. All groups either had English classes at the end of the morning, the third class of the day, or just after lunch, the first class in the afternoon. Those 9-12 year-old students are beginning to be affected by many biological factors during the pre-adolescence period. "We should always bear in mind that different learner profiles and individual differences coexist in a group, so data-driven reflection might not be equally appropriate for all students (Cobb; Boulton, 2015, p. 487).

Once pointing out the areas for further development, it is possible to say that the overall data analysis results indicate that working with KWICs in concordance lines in 4th to 6th grades is possible, is intriguing for younger learners, motivating and can be very effective. As stated by the teacher in her diary, it added variety to the lessons and heightened the level of engagement in some groups. With more DDL work and learners being exposed to specialized authentic vocabulary more frequently from an early age, teachers will be able to notice vocabulary expansion through heightened awareness of language patterns.

We will argue in favor of bringing the concordance lines into the classroom with more pleasurable and motivating DDL work, on different days and with different topics. Or else, DDL approaches can be supplemented, "augmented – or even replaced entirely – with complementary methods of linguistic analysis[63] if DDL is ever going to gain a foothold in mainstream education practice" (Crosthwaite; Boulton, 2023, p. 9, in press). In their words,

[63]See: https://elenlefoll.pressbooks.com.

DDL should evolve "to accommodate ways that learners are used to sourcing information, rather than DDL practitioners trying to force them to adopt KWIC concordancing with limited left / right context (Crosthwaite; Boulton, 2023, p. 9, in press)". In addition, texts which are age and topic-appropriate can turn themselves into a *pedagogic corpus* to be uploaded to the #LancsBox concordancer for hands-on use. The development of the competence in ICT skills will be the ultimate thrust upon learners.

The topics / themes which were dealt with in class showed promising positive short-term retention which can point towards achievement of the main aims. Even though there were no control groups, data for statistical comparisons of outcomes were collected and analyzed, giving researchers a better understanding of the work with pedagogic corpora as it unfolded. Nevertheless, it would have been desirable to administer a second posttest to assess learners' language retention some weeks later, but as we were dealing with young learners, going through some periods of learning recall of previous years' contents after the 2-year of pandemic and lockdown, I decided not to impose further on the teacher.

Despite the positive outcomes, we would name an area where there is much room for further observation and improvement when working with corpora: the selection of target language in pedagogic corpora. To implement this study, both the teacher and researcher analyzed the most frequent vocabulary generated by the concordancer and selected the words and n-grams which we assumed would suit the learners best. However, to our surprise, some of them were not necessarily the ones which called learners' attention. Instead, learners asked for and used words which had not been identified as the most frequent by the software. L1 influence? As mentioned before, many Geography and Science target words worked with in this investigation had not been ranked at the top as the most frequent lexis in the list, something recommended by many studies (McCarten, 2007; Nation, 1997; Nation; Waring, 1997). This fact points out to the need that the class teacher should work collaboratively with the teachers of the subjects involved in future cross-curricular activities. They will be able to help analyze vocabulary from the corpora and curate the most relevant keywords which would be suitable to the needs and interests of the groups. Teachers would then be able to generate the related concordance lines and design the tasks.

There are still other pedagogical implications to be addressed in areas which are beyond our control at present. There is an urgent need that both teachers and learners think realistically about their literacy in ICT as it requires considerable investment in terms of time and practice in order to comprehend the rationale and learn how to use data efficiently (Crosthwaite, 2022; Meunier, 2011, 2022; Boulton, 2012). Meunier (2022) claims that DDL has still not taken

advantage of the numerous advances in digital technology, which in Brazil can also be explained by the lack of computers available to learners in many public schools. The concordance printouts of the 90s have now been upgraded to online DDL systems and there is an urgent need that we close this wide gap as soon as possible. The preparatory courses for teachers still have not included the use of corpora officially. Little digital literacy combined with a reduced number of computers in the majority of public and private schools indicate that class work will probably continue to be hands-off and paper-based in Brazil for quite some time.

Although not meant to be representative of a language in its entirety, the results of this investigation can be particularly useful in teaching English with specific topicalized contents where published materials are difficult to come by (Cobb; Boulton, 2015; Reppen, 2010). Despite the wide array of available English textbooks in Brazil, their target is the general English learners, and do not include the specialized contents of the Elementary schools curricula[64]. The new corpus-informed data can encourage teachers to resort to the pedagogic corpora in cross-curricular projects in English and eventually raise the amount of exposure English learners have in their 4th to 6th grades of *Ensino Fundamental I.* We need to start working with the professional demands of the future now, so the students are prepared to respond to them accordingly when time comes.

After carrying out the study we can reaffirm that there is clear evidence of the effectiveness of DDL in the English classroom (Cobb; Boulton, 2017). Our conclusion is that it is indeed worth trying to introduce models of authentic language – concordance lines and n-grams – in the classroom from the 4th grade or even sooner. Further work on 3- and 4-grams in the concordance lines will also expose learners to mainstream grammar, collocations, and colligations, which are also part of formulaic knowledge that needs to be fostered from their early stages of learning. To corroborate that, O'Keeffe relies on the UB models to argue that "second language learners typically move from a repertoire of fixed holophrasal[65] sequences at low levels to an expanded slot and frame system to fully abstracted (often figurative) patterns (2022, p. 2). Teachers should also resort to the use of computers for more autonomous hands-on discovery work where possible while allowing learners to take agency of their own learning.

---

[64] Corpora will soon be available at CORPUS FOR SCHOOLS – Teaching English Language with Corpus Linguistics website: https://wp.lancs.ac.uk/corpusforschools/. The aim of the project is to bring corpora and corpus methods into classrooms to teach students about the use of the English language.

[65] "In the study of language acquisition, holophrasis is the prelinguistic use of a single word to express a complex idea" (https:en.wikipedia.org/wiki/Holophrasis) and "this is often observed in language acquisition, where children in the prelinguistic stage use holophrases to communicate complex ideas" (https://thoughtco.com/holophrase-language-acquisition-1690929).

It is fair to say that if there is a continuation of classwork with corpus-informed data and a DDL approach, language will be consolidated.

We conclude endorsing Boulton's words about his own work, "the simple experiment [...] is deliberately modest in its design and aims in order to show that useful empirical results are not hard to obtain, and in the hope that others may therefore be encouraged to conduct their own empirical studies" (Boulton, 2008, p. 1). We did!

References

ALEXANDER, Richard John. Fixed expressions in English: a linguistic, psycholinguistic, sociolinguistic and didactic study. *Anglistik und Englischunterricht*, v. 6, n. 1978, p. 171-188, 1979.

ALMEIDA, Valdênia; ORFANÓ, Bárbara; DUTRA, Deise. Is there a better choice? Verb-noun combinations in academic writing. *In*: VIANA, Vander (org.). *Teaching English with corpora:* a resource book. Oxon: Routledge, 2023. p. 228-232.

ANTHONY, Laurence; CHUJO, Kiyomi; OGHIGIAN, Kathryn . *DDL for the EFL classroom* - Effective uses of a Japanese-English parallel corpus and the development of a learner-friendly, online parallel concordancer. Waseda University, 2009.
DOI: https://doi.org/10.1163/9789401206884_008 .

ANTHONY, Laurence; CHUJO, Kiyomi; OGHIGIAN, Kathryn; YOKOTA, Kenji. Teaching Remedial Grammar through Data-Driven learning using AntPConc. *Taiwan International ESP Journal*, v. 5, n. 2, p. 65-90, 2013.

ARMSTRONG, Patricia. *Bloom's Taxonomy*. Vanderbilt University Center for Teaching, 2010. Available at: https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy. Accessed on July 1st., 2023.

ASTON, Guy. Small and large corpora in language learning. *In*: LEWANDOWSKA-TOMASZCZYK, Barbara; MELIA, Patrick James (eds.). PALC'97. *Proceedings of the first annual conference*. Łodz: Łodz University Press, 1997. p. 51-62.

ASTON, Guy. Corpora in Language Pedagogy: matching theory and practice. In: COOK, Guy; SEILDHOFER, Barbara (Eds). Principle and practice in applied linguistics: studies in honour of H. G. Widdowson. Oxford: Oxford University Press, 1995. p. 385 – 418.

BANERJEE, Satanjeev; PEDERSEN, Ted. Extended Gloss Overlaps as a Measure of Semantic Relatedness. INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE – IJCAI, 18th, Acapulco, Mexico, 2003. *Proceedings…* Morgan Kaufmann Publishers Inc. San Francisco, CA, United States.

BIBER, Douglas; CONRAD, Susan; LEECH, Geofrey. *Longman Grammar of Spoken and Written English*. Essex: Pearson Education, 1999.

BIBER, Douglas; CONRAD, Susan; REPPEN, Randi. *Corpus Linguistics.* Investigating language structure and use. Cambridge: Cambridge University Press, 1998.

BINGIMLAS, Khalid. Barriers to the successful integration of ICT in teaching and learning environments: A review of the literature. *Eurasia journal of mathematics, science & technology education*, v5., n. 3, 2009. Available at: https://doi.org/10.12973/ejmste/75275. Accessed on Aug. 1st., 2022.

BLOOM, Benjamin; ENGLEHART, Max; FURST, Edward; HILL, Walter; KRATHWOHL, David. *The Taxonomy of educational objectives, Handbook I*: The Cognitive domain. New York: David McKay Co. Inc., 1956.

BOCORNY, Ana Elisa; REBECHI, Rozane; REPPEN, Randi; DELFINO, Maria Cláudia; LAMEIRA, Vivian. A produção de artigos da área das ciências da saúde com o auxílio de key lexical bundles: um estudo direcionado por corpus. *D.E.L.T.A*. [Documentação de Estudos em Linguística Teórica e Aplicada], v. 37, p. 1, 2021. Available at: https://doi.org/10.1590/1678-460X2021370101. Accessed on Aug. 2nd, 2021.

BOCORNY, Ana Elisa; WELP, Anamaria. O desenho de tarefas pedagógicas para o ensino de inglês para fins acadêmicos: conquistas e desafios da Linguística de Corpus [The design of pedagogical tasks for teaching English for academic purposes: achievements and challenges of Corpus Linguistics]. *Revista de Estudos da Linguagem*, v. 29, p. 1529, 2021. Available at: http://dx.doi.org/10.17851/2237-2083.29.2.1529-1638. Accessed on Aug. 1st., 2021.

BOLINGER, Dwight. Meaning and memory. *Forum Linguisticum*, v. 1, p. 1-14, 1976.

BOULTON, Alex. Research in data-driven learning. *In*: PÉREZ-PAREDES, Pascual; MARK, Geraldine (eds.). *Beyond Concordance Lines*: Corpora in language education. John Benjamins, 2021. p. 9-34.

BOULTON, Alex. Foreword. *Data-Driven Learning for the Next Generation*. New York: Routledge, 2020. [e-book – Kindle Edition].

BOULTON, Alex. What data for data-driven learning? *Eurocall Rev.*, v. 20, 2012a. p. 23–27.

BOULTON, Alex. Hands-on / hands-off: Alternative approaches to data-driven learning. *In*: THOMAS, James; BOULTON, Alex (eds.). *Input, Process and Product*: Developments in Teaching and Language Corpora. Brno, Czech Republic: Masaryk University Press, 2012. p. 152-168.

BOULTON, Alex. Looking for empirical evidence of data-driven learning at lower-levels. *In*: LEWANDOSKA-TOMASZCZYK, Barbara (ed.). *Corpus Linguistics, Computer Tools, and Applications: State of the Art.* Frankfurt: Peter Lang, 2008. p. 581-598.

BRASIL. Parecer CNE/CEB n.01/2020. *Diretrizes curriculares nacionais para a educação bilíngue* [White paper CNE/CEB - National curriculum guidelines for bilingual education]. Brasília: Ministry of Education. Available at: http://portal.mec.gov.br/docman/maio-2020-pdf/146571-texto-referencia-parecer-sobre-educac-a-o-bili-ngue/file. Accessed on June 1st., 2020.

BRASIL. Base Nacional Comum Curricular (BNCC). Ministério da Educação e Cultura. Disponível em: https://novaescola.org.br/conteudo/12720/bncc-baixe-em-pdf-o-e-book-de-competencias-gerais. Acesso em: 1 abr. 2021.

BRASIL. Ministério da Educação; Secretaria de Educação Básica; Secretaria de Educação Continuada, Alfabetização, Diversidade e Inclusão; Secretaria de Educação Profissional e Tecnológica. Conselho Nacional de Educação; Câmara de Educação Básica. *Diretrizes*

Curriculares Nacionais da Educação Básica. Brasília: MEC; SEB; DICEI, 2013. Disponível em: http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=13448-diretrizes-curiculares-nacionais-2013-pdf&Itemid=30192. Acesso em: 1 abr. 2021.

BRASIL. Ministério da Educação. Conselho Nacional da Educação. *Base Nacional Comum Curricular*, 2017. Brasília, Secretaria de Educação Básica. Disponível em: http://basenacionalcomum.mec.gov.br/images/BNCC_EI_EF_110518_versaofinal_site.pdf. Acesso em: 1 abr. 2021.

BREZINA, Vaclav; TIMPERLEY, Mathew; MCENERY, Anthony. *#LancsBox* v.4.x. [software]. Lancaster University, 2018. Available at: http://corpora.lancs.ac.uk/lancsbox/download.php. Accessed on Dec. 1st., 2019. [#LANCSBOX]

BREZINA, Vaclav; PLATT, William. *#LancsBox*: corpus toolbox v.6.0. [software]. Lancaster University, 2021. Available at: #LancsBox: Lancaster University corpus toolbox. Accessed: Dec. 1st. 2021. [#LANCSBOX]

BROOKS, Nelson. *Language and language learning*: Theory and practice (2nd. ed.). New York: Harcourt Brace & World, 1964.

CAMPBELL, Donald; STANLEY, Julian. *Experimental and quasi-experimental designs for research*. Palo Alto, CA: Houghton Mifflin Company Boston, 1966.

BROWN, Roger. *A first language*: The early stages. Cambridge, MA: Harvard University Press, 1973.

CARTER, Ronald. *Vocabulary*. London: Routledge, 1987.

CHAMBERS, Angela. Towards the corpus revolution? Bridging the research – practice gap. *Language Teaching*, v. 52, n. 4, p. 460-475, 2019.

CHAMBERS, Angela. What is data-driven learning? *In*: O'KEEFFE, Anne; MCCARTHY, Michael (eds.). *The Routledge Handbook of Corpus Linguistics.* London: Routledge, 2010. p. 345-358.

CHAMBERS, Angela; KELLY, Victoria. Semi-specialized corpora of written French as a resource in language teaching and learning. *TEANGA,* v. 21, p. 20-21, 2002.

CLEMESHA, Susan; LIBERALI, Fernanda. De-encapsulated Bilingual Education in Brazil Multicultural Breakfasts and Translanguaging Kids. *ReVista: Harvard Review of Latin America*, v. XIX, n. 2, Fall/Winter 2019-2020. Available at: https://revista.drclas.harvard.edu/de-encapsulated-bilingual-education-in-brazil/. Accessed on July 1st., 2022.

COBB, Thomas. *Compleat Lexical Tutor* v. 8.3 – *VP-Kids v. 1.1.* Available at: https://www.lextutor.ca/vp/kids/. Accessed on April 1st., 2021.

COBB, Thomas; BOULTON, Alex. Corpus use in language learning: A meta-analysis. *Language Learning*, Wiley, v. 67, n. 2, p. 348 – 393, 2017.

COBB, Thomas; BOULTON, Alex. Classroom Application of Corpus Analysis. *In*: BIBER, Douglas; REPPEN, Randi (eds.). *The Cambridge Handbook of English Corpus Linguistics.* Cambridge: Cambridge University Press, 2015. p. 480-497.

COBB, Thomas. Computing the Vocabulary Demands of L2 Reading. *In*: CHUN, Dorothy; HEIFT, Trude (eds.). *Language Learning and Technology*, v.11, n. 3, 2007. p. 38-63. Available at: http://llt.msu.edu/vol11num3/cobb/. Accessed on April 1st., 2021.

COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES. Available at: https://www.cambridgeenglish.org/images/177867-the-methodology-behind-the-cambridge-english-scale.pdf. Accessed on April 1st., 2021.

CORINO, Elisa; ONESTI, Cristina. Data-Driven Learning: A Scaffolding Methodology for CLIL and LSP Teaching and Learning. *Frontiers in Education*. 2019. Available at: https://doi.org/10.3389/feduc.2019.00007. Accessed on Aug. 1st, 2022.

CORTES, Viviana. Lexical bundles in published and student disciplinary writing: Examples from History and Biology. *English for Specific Purposes*, v. 23, n. 4, p. 397-423, 2004.

COXHEAD, Averil. New academic word list. *TESOL Quarter*, v. 34, p. 213-238, 2000.

COYLE, Do; HOOD, Philip; MARSH, David. *Content and Language Integrated Learning.* Cambridge: CUP, 2010.

CROSTHWAITE, Peter; BOULTON, Alex. DDL is dead? Long live DDL! Expanding the boundaries of data-driven learning. *In*: TYNE, H.; BILGER, M.; BUSCAIL, L; LERAY, M.; CURRY, N.; PÉREZ-SABATER, C. (eds.). *Discovering language*: Learning and affordance. Peter Lang, 2023. [in press].

CROSTHWAITE, Peter; BAISA, Vit. Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, v. 3, n. 3, p.1-4, 2023. Available at: https://doi.org/10.1016/j.acorp.2023.100066.

CROSTHWAITE, Peter. DDL for Younger Learners. *In*: JABLONKAI, Reka; CSOMAY, Eniko (eds.). *The Routledge Handbook of Corpora and English Language Teaching and Learning*. New York: Routledge, 2022. p. 377-389.

CROSTHWAITE, Peter; STELL, Annita. It helps me get ideas on how to use my words – Primary school students' initial reactions to corpus use in a private tutoring setting. *In*: CROSTHWAITE, Peter (ed.). *Data-Driven Learning for the Next Generation* - Corpora and DDL for Pre-tertiary Learners. New York: Routledge, 2020. [e-book – Kindle Edition, Locations: 3837-3838].

DANG, Ngoc Yen; WEBB, Stuart. Evaluating lists of high-frequency words. *International Journal of Applied Linguistics*, v. 167, n. 2, p. 132-158, 2016.

DANG, Thi Ngoc Yen; WEBB, Stuart. Making an essential word list for beginners. *In*: NATION, Paul (ed.). *Making and using word lists for language learning and testing*. Amsterdam: John Benjamins, 2016a. p. 153-167, 188-195.

DE CHIARA, Márcia; GAVRAS, Douglas. Desempregado sem formação não consegue nem trabalhos básicos: mutirão de emprego não preenche 60% das vagas ofertadas. *O Estado de São Paulo*, São Paulo, 2 jun. 2019. Capa e Suplemento B – Economia, p. 4.

DE OLIVEIRA, Luciana; HÖFLING, Camila. Bilingual Education in Brazil. *In*: RAZA, Kashif; COOMBE, Christine; REYNOLDS, Dudley (eds.). *Policy Development in TESOL and Multilingualism* – Past, Present and the Way Forward. Springer, 2021. p. 5-37. DOI:10.1007/978-981-16-3603-5.

DOLCH, Edward. A basic sight word vocabulary. *The Elementary School Journal*, v. 36, n. 6, p. 456-460, 1936.

DUTRA, Deise; QUEIROZ, Jessica; de MACEDO, Luciana; COSTA, Danilo; MATTOS, Elisa. Adjectives as nominal pre-modifiers in chemistry and applied linguistics research articles. *In*: RÖMER, Ute; CORTES, Viviana; FRIGINAL, Eric (org.). *Advances in Corpus-based Research on Academic Writing*. Amsterdam: John Benjamins Publishing Company, 2020. DOI: https://doi.org/10.1075/scl.95.09dut.

ELLIS, Nick. Frequency effects. *In*: ROBINSON, Peter. (ed.) *The Routledge Encyclopedia of Second Language Acquisition*. New York: Routledge, 2012. p. 260-265.

ELLIS, Nick; LARSEN-FREEMAN, Diane. Language emergence: implications for applied linguistics. Introduction to the special issue. *Applied Linguistics*, v. 27, n. 4, p. 558-589, 2006.

ELLIS, Nick. Frequency effects in Language Processing – A Review with implications for Theories of Implicit and Explicit language Acquisition. *Studies in Second Language Acquisition,* Cambridge, v. 24, n. 2, 2002. p. 143-188. DOI: https://doi.org/10.1017/S0272263102002024.

ELLIS, Nick. Memory for Language. *In*: ROBINSON, Peter (ed.). *Cognition and Second Language Instruction.* Cambridge: Cambridge University Press, 2001. p. 33-68.

ELLWOOD, Nancy (ed.). *Geography – 3rd Grade*. New York: DK – Penguin Random House, 2014.

EHRI, Linnea. Learning to Read Words: Theory, Findings, and Issues. *Scientific Studies of Reading*, v. 9, n. 2, p. 167-188, 2005.

EHRI, Linnea. Reconceptualizing the development of sight word reading and its relationship to recoding. In: GOUGH, Philip *et al* (eds.). *Reading Acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc, 1992. p. 107–143.

FARRELL, Linda; OSENGA, Tina; HUNTER, Michael. Comparing the Dolch and Fry High Frequency Word Lists. *Readsters*, 2013. p. 1-14. Available at: https://www.readsters.com/wp-content/uploads/2013/03/ComparingDolchAndFryLists.pdf. Accessed on: April 20, 2021.

FLOWERDEW, Lynne. Data-driven learning and language learning theories: whither the twain shall meet. *In*: LENKO-SZYMANSKA, Agnieszka; BOULTON, Alex (eds.). *Multiple Affordances of Language Corpora for Data-Driven Learning*. Amsterdam, Netherland: John Benjamins, 2015. p. 15-36.

FORSBERG LUNDELL, Fanny. Formulaicity and corpora. *In*: TRACY-VENTURA, Nicole; PAQUOT, Magali (eds.). *The Routledge Handbook of Second Language Acquisition and Corpora.* Routledge: London, 2021. p. 370-381.

FRANKENBERG-GARCIA, Ana. Raising teachers' awareness of corpora. *Language Teaching*, v. 45, n. 4, p. 475-489, 2012.

FRIES, Charles; LADO, Robert. *Teaching Pronunciation*. Michigan: University of Michigan Press, John Wiley, 1958.

FRY, Edward. The new instant word list. *The Reading Teacher,* v. 34, n. 3, p. 284-289, 1980.

GABLASOVA, Dana. *Corpus Linguistics: Method, Analysis and Interpretation.* MOOC. Lancaster: Lancaster University, 2019.

GABRIELATOS, Costas. Corpora and Language Teaching: Just a fling or wedding bells? *TESL – EJ*, v. 8, n. 4, p. 1-37, 2005.

GARNER, Jeremy. *Vocabulary Kitchen*. 2019. Available at: http://vocabkitchen.com/profiler/cefr. Accessed on: April 2nd, 2021.

GIBBONS, Pauline. *Scaffolding Language, Scaffolding Learning*: Teaching Second Language Learners in the Mainstream Classroom. Sydney: Heinemann, 2002.

GILQUIN, Gaetanelle; GRANGER, Sylviane. How can data-driven learning be used in language teaching? *In*: O'KEEFFE, Anne; MCCARTHY, Michael (eds.). *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 2010. p. 359-371.

GODWIN-JONES, Robert. The evolving roles of language teachers: trained coders, local researchers, global citizens. *Language Learning & Technology*, v. 19, n. 1, p. 10-22, 2015.

GRANGER, Sylviane. The computer learner corpus: a versatile new source of data for SLA research. *In*: GRANGER, S. (ed.). *Learner English on Computer*. London, New York: Addison Wesley Longman, 1998. p. 3-18.

HAFNER, Christoph; CANDLIN, Christopher. Corpus tools as an affordance to learning in professional legal education. *Journal of English for Academic Purposes*, v. 6, n. 4, p. 303-318, 2007.

HARLEY, Birgit; WANG, Wenxia. The critical period hypothesis: where are we now? *In*: DE GROOT, Annette M. B.; KROLL, Judith F. (eds.). *Tutorials in Bilingualism*: Psycholinguistic Perspectives. Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers, 1997. p. 19-51.

HARLEY, Birgit; WANG, Wenxia. Age and the critical period hypothesis. *ELT Journal*, v. 63, n. 2, p. 170-172, 2009. DOI: 10.1093/elt/ccn072.

HARMER, Jeremy. *How to Teach English*. London: Longman, 2003.

HENDRY, Clinton; SHEEPY, Emily. Evaluating corpus analysis tools for the classroom. *In*: JABLONKAI, Reka; CSOMAY, Eniko (eds.). *The Routledge Handbook of Corpora and English Language Teaching and Learning*. New York: Routledge, 2022. p. 437-459.

HIRATA, Eri. The development of a multimodal corpus for young learners: a case study on the integration of DDL in teacher. *In*: CROSTHWAITE, Peter (ed.). *Data-Driven Learning for the Next Generation* - Corpora and DDL for Pre-tertiary Learners. New York: Routledge, 2020. [e-book – Kindle Edition, Locations: 2486 - 2895].

HIRSH, David; NATION, Paul. What Vocabulary Size is Needed to Read Unsimplified Texts for Pleasure? *In*: DAY, Richard; BRANTMEIER, Cindy (eds.). *Reading in a Foreign Language*, v. 8, n. 2, p. 689-696, 1992.

IBM CORP. SPSS Statistics for Windows. IBM CORP. Released in: 2012. [S.I.]: IBM SPSS INC., 2012. Available at: https://www.ibm.com/products/spss-statistics. Accessed on: August 1st., 2022.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO). Available at: https://www.iso.org/home.html. Accessed on: May 1st., 2023

JOHANSSON, Stig. Some thoughts on corpora and second-language acquisition. *In*: AIJMER, Karin. (ed.). *Corpora and Language Teaching*. Amsterdam: John Benjamins, 2009. p. 33-44.

JOHNS, Tim. Contexts: the background, development and trialing of a concordance-based CALL program. *In*: WICHMANN, Anne *et al.* (eds.). *Teaching and Language Corpora*. Harlow: Longman, 1997. p. 100-115.

JOHNS, Tim. Should you be persuaded: Two examples of data driven learning. *In*: JOHNS, Tim; KING, Philip (eds.). Classroom concordancing. *ELR Journal*, v. 4. Birmingham: University of Birmingham, 1991. p. 1-16.

JOHNS, Tim. Micro-concord: a language-learner's research tool. *System*, v. 14, n. 2, p. 151-162, 1986.

JUDD, Elliot; TAN, Libua; WALBERG, Herbert. *Teaching additional languages*. New York: International Academy of Education / UNESCO, 2001.

LABERGE, David; SAMUELS, Jay. Toward a theory of automatic information processing in reading. *Cognitive Psychology*, v. 6, p. 293-323, 1974.

LANGACKER, Ronald. *Foundations of Cognitive Grammar*: Theoretical Prerequisites. Stanford: Stanford University Press, 1987.

LANCASTER University corpus toolbox. Version 6.0. Available at: #LancsBox: Lancaster University corpus toolbox. Accessed on: August 1st., 2021. [#LANCSBOX]

LAUFER, Batia; NATION, Paul. Vocabulary. *In*: GASS, Susan; MACKEY, Alison (eds.). *The Routledge Handbook of Second Language Acquisition*. New York: Routledge, 2012. p. 163-176.

LEEL, Hsing-chin. In Defense of Concordancing: An Application of Data-Driven Learning in Taiwan. *Procedia - Social and Behavioral Sciences*, v. 12, p. 399-411, 2011.

LEWIS, Michael. *Implementing the Lexical Approach.* Hove, UK: LTP – Language Teaching Publications, 1993.

LIBERALI, Fernanda. A BNCC e a elaboração de currículos para a Educação Bilíngue. *In*: MEGALE, Antonieta. (ed.). *Educação Bilíngue no Brasil*. São Paulo: Fundação Santillana, 2019. p. 29-42.

LIGHTBOWN, Patsy; SPADA, Nina. *How Languages are Learned*. Oxford: Oxford University Press, 2013.

LONG, Michael. Focus on Form: A Design feature in Language Teaching. *In*: DE BOT, Kees *et al.* (eds.). *Foreign Language Research in Cross-Cultural Perspectives*. Amsterdam: John Benjamins, 1991. p. 39-52.

MAIA, Belinda. Do-it-yourself corpora… with a little help from your friends. *In*: LEWANDOWSKA-TOMASZCZYK, Barbara; MELIA, Patrick James (eds.). PALC '97 *Proceedings of the first annual conference.* Łodz: Łodz University Press, 1997. p. 403-410.

MARINOVA-TODD, Stefka; MARSHALL, D. Bradford; SNOW, Catherine. Three Misconceptions about Age and L2 Learning. *TESOL Quarterly*, v. 34, n. 1, p. 9-34, Spring 2000. Available at: https://www.jstor.org/stable/3587863. Accessed on: April 2nd. 2021.

MARSH, David. Bilingual Education & Content and Language Integrated Learning. *International Association for Cross-cultural Communication, Language Teaching in the Member States of the European Union (Lingua)*, Paris, University of Sorbonne, 1994.

MCCARTEN, Jeanne. *Teaching vocabulary.* Lessons from the Corpus, Lessons for the Classroom. Cambridge: Cambridge University Press, 2007.

MCCARTHY, Michael. *Touchstone*: from Corpus to Coursebook. Cambridge: Cambridge University Press, 2004.

MCENERY, Tony; WILSON, Andrew. *Corpus Linguistics*. 2nd ed. Edinburgh, UK: Edinburgh University Press, 2001.

MCKENNEY, Susan; VISSCHER, Adrie. Technology for Teacher Learning and Performance. *Technology, Pedagogy and Education,* v. 28, p. 129-132, 2019. DOI: https://doi.org/10.1080/1475939X.2019.1600859. Accessed on: August 2nd. 2021.

MEARA, Paul. *The importance of an Early Emphasis on L2 Vocabulary*. University of Wales, Swansea, 2001. Available at: http://www.jalt-publications.org/old_tlt/files/95/feb/meara.html. Accessed on: April 3rd , 2021.

MEARA, Paul. Single-subject studies of lexical acquisition. *Second Language Research*. Swansea, University of Wales, v. 11, n. 2, p. i-iii, 1995.

MELKA, Francine. Receptive vs. productive aspects of vocabulary. In: SCHMITT, Norbert; MCCARTHY, Michael (eds.). *Vocabulary*: Description, Acquisition and Pedagogy. Cambridge: Cambridge University Press, 1997. p. 84–102.

MENDES, Ana Rachel; FINARDI, Kyria. Integrating digital technologies in Brazilian English language Teacher Education through blended learning. *Educação em Revista,* v. 36, e. 233799, 2020. DOI: http://dx.doi.org/10.1590/0102-4698233799. Accessed on: Aug. 3rd , 2022.
MEUNIER, Fanny. Revamping DDL: Affordances of Digital Technology. *In*: JABLONKAI, Reka; CSOMEY, Eniko (eds.). *The Routledge Handbook of Corpora and English Language Teaching and Learning.* New York, London: Routledge, 2022. p. 344-360.

MEUNIER, Fanny. A Case for Constructive Alignment in DDL: Rethinking Outcomes, Practices and Assessment in (Data-Driven) Language Learning. *In*: CROSTHWAITE, Peter (Ed.). *Data-Driven Learning for the Next Generation*. London: Routledge, 2020. p. 13-30. DOI: https://doi.org/10.4324/9780429425899-2.

MEUNIER, Fanny; REPPEN, Randi. Corpus versus non-corpus-informed pedagogical materials: Grammar as the focus. *In*: BIBER, Douglas; REPPEN, Randi (eds.). *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, 2015. p. 498-514.

MEUNIER, Fanny. Corpus linguistics and second / foreign language learning: exploring multiple paths. *Revista Brasileira de Linguística Aplicada*, Belo Horizonte, v. 11, n. 2, p. 459-477, 2011.

NATION, Paul. What matters in vocabulary learning? *LALS*, Victoria University of Wellington, New Zealand, 2020. [Webinar].

NATION, Paul. *What should every ESL teacher know?* Seoul: Compass Publishing, 2013.

NATION, Paul. *Learning vocabulary in another language*. 2nd. ed. Cambridge: Cambridge University Press, 2013a.

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

NATION, Paul. The goals of vocabulary learning. *In*: NATION, Paul (ed.). *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press, 2001. p. 6-22.

NATION, Paul. *Teaching and Learning Vocabulary.* New York: Heinle and Heinle, 1990.

NATION, Paul; WARING, Robert. Vocabulary size, text coverage and word lists. *In*: SCHMITT, Norbert; MCCARTHY, Michael (eds.). *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press, 1997. p. 6-19.

O'KEEFFE, Anne; MARK, Geraldine. Principled pattern curation to guide data-driven learning design. *Applied Corpus Linguistics*. 2022. DOI: https://doi.org/10.1016/j.acorp.2022.100028. Accessed on: Sept. 1st. 2022.

O'KEEFFE, Anne. Data-driven learning and the second language acquisition interface debate. *In*: PÉREZ-PAREDES, Pascual; MARK, Geraldine (eds.*). Beyond the Concordance*: Multiple Applications of Language Corpora for Language Education. Amsterdam: John Benjamins, 2021a. p. 35-55.

O'KEEFFE, Anne. Data-driven learning – a call for a broader research gaze. *Language Teaching,* v. 54, n. 2, p. 259-272, 2021b.

O'KEEFFE, Anne; MARK, Geraldine. The English grammar profile of learner competence: methodology and key findings. *International Journal of Corpus Linguist*, v. 22, n. 4, p. 457-489, 2017.

O'KEEFFE, Anne; MCCARTHY, Michael; CARTER, Ronald. *From corpus to classroom*: language use and language teaching. Cambridge, UK: Cambridge University Press, 2007.

OLIVA, Katherine. *Integrating corpora with language classroom*: online corpus as an editing tool. Orientadora: Dra. Deise Prina Dutra. 117 f. 2018. Dissertação (Mestrado em Linguística Aplicada) - Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2018.

OROSZ, Andrea. The growth of young learners' English vocabulary size. *In*: NIKOLOV, Marianne (ed.). *Early learning of modern foreign languages*. Bristol: Multilingual Matters, 2009. p. 181-194.

PÉREZ-PAREDES, Pascual. The pedagogic advantage of teenage corpora for secondary school learners. *In*: CROSTHWAITE, Peter (ed.). *Data-Driven Learning for the Next Generation.* New York: Routledge, 2020. [Kindle Edition. Loc. 2021-2482].

PINTER, Annamaria. Teaching Young Learners. *In*: BURNS, Anne; RICHARDS, Jack (eds.). *The Cambridge Guide to Pedagogy and Practice in Second Language Teaching.* Cambridge: Cambridge University Press, 2012. p. 103-111.

PINTO, Paula; CROSTHWAITE, Peter; de CARVALHO, Carolina; SPINELLI, Francieli; SERPA, Talita; GARCIA, William; OTTAIANO, Ariane (orgs.). *Using language data to Learn About Language*: A Teacher's Guide to Classroom Corpus Use. Brisbane: University of Queensland, 2023. DOI:10.14264/3bbe92d.

RAYSON, Paul. Computational tools and methods for corpus compilation and analysis. *In*: BIBER, Douglas; REPPEN, Randi (eds.). *The Cambridge Handbook of English Corpus Linguistics.* Cambridge: Cambridge University Press, 2015. p. 32-49.

READ, Carol. *ABC of Teaching Children – ideas, tips and resources for primary language teachers.* Available at: https://carolread.wordpress.com/2011/07/25/y-is-for-young-learners/. Accessed on: May 1st. 2021.

READ, Carol. Is younger better? *English Teaching professional*, v. 28, p. 5-7, July, 2003.

READ, Carol. ABC of changes in PELT over the last 30 years. *IATEFL Young Learners & Teenagers SIG Newsletter*: C&TS Digital Special Pearl Anniversary Edition. Available at: https://www.carolread.com/wp-content/uploads/2015/01/ABC-of-changes-in-PELT-over-the-last-30-years.pdf. Accessed on: November 1st 2023.

REDECKER, Christine. European Framework for the Digital Competence of Educators: DigCompEdu. *JRC Research Reports*, JRC107466, Joint Research Centre, Seville, European Commission, 2017.

REPPEN, Randi. Using Corpora in the Language Classroom, 2012. (1:17:32). Published by: The New School. Available at: https://www.youtube.com/watch?v=Qf46lOnMCfs&t=967s. Accessed on: August 4th 2020.
REPPEN, Randi. *Using Corpora in the Language Classroom.* Cambridge: Cambridge University Press, 2010.

REPPEN, Randi. Building a Corpus – What are the key considerations? *In*: O'KEEFFE, Anne; MCCARTHY, Michael (eds.). *Routledge Handbook of Corpus Linguistics*. London: Routledge, 2010a. p. 31-37.

RIBEIRO, Elisa. Que futuros redesenhamos? Uma releitura do manifesto da Pedagogia dos Multiletramentos e seus ecos no Brasil para o século XXI. *Diálogo das Letras*, Pau dos Ferros, v. 9, p. 1-19, 2020.

RICHARDS, Jack. Preface. *In*: REPPEN, Randi. *Using Corpora in the Language Classroom.* Cambridge: Cambridge University Press, 2010.

RICHARDS, Jack; RODGERS, Theodore. *Approaches and Methods in Language Teaching.* Cambridge: Cambridge University Press, 1986; 2014.

RÖMER, Ute. A corpus perspective on the development of verb constructions in second language learners. *International Journal Corpus of Linguistics*, v. 24, n. 3, 2019, p. 268-290.

RUBIO, Fernando; PASSEY, Amber; CAMPBELL, Selene. Grammar in disguise: the hidden agenda of communicative language teaching textbooks. *RAEL*: revista electrónica de linguística aplicada, n. 3, p. 158-176, 2004. ISSN 1885-9089. Available at: https://www.researchgate.net/publication/28102295_Grammar_in_disguise_the_hidden_agenda_of_communicative_language_teaching_textbooks. Accessed on: May 5th 2022.

RUTHERFORD, William. *Second language grammar*: Learning and teaching. London: Longman, 1987.

SARDINHA, Tony Berber; ZUPPARDI, Maria Carolina. A Multi-dimension view of collocations in academic writing. *In*: RÖMER, Ute; CORTES, Viviana; FRIGINAL, Eric (orgs.). *Advances in Corpus-based Research on Academic Writing*: Effects of discipline, register and writer expertise. John Benjamins Publishing Company, 2020. p. 333-354.

SCHIMDT, Richard. The role of consciousness in second language learning. *Applied Linguistics*, v. 11, n. 2, p. 129-158, 1990.

SCHMITT, Norbert; SCHMITT, Diane; CLAPHAM, Caroline. Developing and exploring the behavior of two new versions of the Vocabulary Levels Test. *Language Testing*, v. 18, n. 1, p. 55-88, 2001.

SHARWOOD SMITH, Mike. Input enhancement in instructed SLA. *Studies in Second Language Acquisition*, v. 15, n. 2, p. 165–179, 1993.
SINCLAIR, John. *Trust the text* – Language, corpus and discourse. London: Routledge, 2004.

SINCLAIR, John. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.

SKINNER, Burrhus. *Verbal behavior*. New York: Appleton-Century-Crofts, 1957.

SEALEY, Alison. The use of corpus-based approaches in children's knowledge about language. *In*: ELLIS, Sue; MCCARTNEY, Elspeth (eds.). *Applied Linguistics and Primary School Teaching*. Cambridge University Press, 2011. p. 93-106.

SOLVES, Tony. Introduction to CLIL. *Laboratorio de Idiomas*, Universitas Miguel Hernandéz, 2018.

STEMACH, Jerry; WILLIAMS, William. *WordExpress*: The first 2,500 words of spoken English. Novato, CA: Academic Therapy Publications, 1988.

SWEET, Henry. *The practical study of languages*: A guide for teachers and learners. New York: Henry Holt and Company, 1899.

TARTONI, Marlei Rose. *A linguística de corpus no ensino de inglês*: um experimento com atividades com "to" e "for" como keywords. Orientadora: Profa. Dra. Deise Prina Dutra. 2012. 164 f. Dissertação (Mestrado em Linguística Aplicada) - Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2012.

TINKHAM, Thomas. The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second Language Research*, v. 13, n. 2, p.138–163, 1997.

TONO, Yukio, DÍEZ-BEDMAR, Maria Belén. Focus on learner writing at the beginning and intermediate stages: the ICCI corpus. *International Journal of Corpus Linguistics*, v. 19, n. 2, p. 163–177, 2014.

THORNDIKE, Edward. *The Teacher's Word Book.* 1st ed. New York: Teacher's College, Columbia University, 1921. Available at: https://archive.org/details/teacherswordbook00thoruoft/page/n7. Accessed on: April 2nd 2021.

THORNDIKE, Edward. *The Teacher's Word Book.* 2nd ed. New York: Teacher's College, Columbia University, 1927. Available at: https://archive.org/details/cu31924014451409/page/n7. Accessed on: April 2nd 2021.

TRIBBLE, Christopher. Put a corpus in your classroom: Using a computer in vocabulary development. *In*: BOSWOOD, Tim (ed.). New ways of using computers in language teaching. *TESOL*, Alexandria, VA, p. 266-268, 1997.

TRIBBLE, Christopher. Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching. FIRST INTERNATIONAL CONFERENCE – PRACTICAL APPLICATIONS IN LANGUAGE CORPORA, 1st., 1997. *Proceedings* [...] University of Łodz, Poland. 1997a. Available at: https://www.ctribble.co.uk/text/Tribble_C_Palc_97.pdf. Accessed on: April 2nd 2021.

VYATKINA, Nina. Corpora as open educational resources for language teaching. *Foreign Language Annals*, v. 53, n. 2, p. 359-370, 2020. Available at: https://doi.org/10.1111/flan.12464. Accessed on: May 1st 2022.

VYGOTSKY, Lev. *Thinking and Speaking*. Cambridge, MA: MIT Press, 1962.

WARD, Jeremy. How large a vocabulary do EAP engineering students need? *Reading in a Foreign Language*, v. 12, n. 2, p. 309-324, 1999.

WEBB, Stuart; NATION, Paul. *How Vocabulary is Learned*. Oxford: Oxford University Press, 2017.

WEBB, Stuart. Depth of vocabulary knowledge. *In*: CHAPPELLE, Carol (ed.). *Encyclopedia of Applied Linguistics.* Oxford, UK: Wiley-Blackwell, 2013. p. 1656–1663.

WEBB, Stuart; CHANG, Anna. Second language vocabulary growth. *RELC Journal*, v. 43, n. 1, p. 113-126, 2012.

WEBB, Stuart; CHANG, Anna. Vocabulary learning through assisted and unassisted repeated reading. *Canadian Modern Language Review*, v. 68, n. 3, p. 1-24, 2012a.

WICHER, Oliver. Data-driven learning in the secondary classroom: a critical evaluation from the perspective of foreign language didactics. CROSTHWAITE, Peter (ed.). *In*: *Data-driven learning for the next generation*: Corpora and DDL for pre-tertiary learners. London: Routledge, 2020. p. 31-46.

WILLIS, Dave. The language syllabus: building language study into a task-based approach. *Classroom Matters*, v. 30, Spring 2011. Available at: http://ihjournal.com/the-language-syllabus-building-languag-into-a-task-based-approach-by-dave-willis-2e-study. Accessed on: April 1st 2021.

WILLIS, Dave. The Language Syllabus: Why Not Start With Lexis? *In: Classroom Matters*, v. 29, Autumn 2010. Available at: http://ihjournal.com/the-language-syllabus-why-not-start-with-lexis-by-dave-willis. Accessed on: April 2nd 2021.

WILLIS, Dave. *The Lexical Syllabus*. London: Collins ELT, 1990.

WILLIS, Jane. Concordances in the Classroom without a computer: Assembling and exploiting Concordances of Common Words. *In*: TOMLINSON, Brian (ed.). *Materials Development in Language Teaching.* Cambridge: Cambridge University Press, 1998. p. 44-66.

WINNE, Philip. Cognition and metacognition within self-regulated learning. *In*: SCHUNK, Dale; GREENE, Jeffrey (eds.). *Routledge Handbook of Self-Regulation of Learning and Performance*. London: Routledge, 2017. p. 36-48.

WRAY, Alison. What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics,* v. 32, p. 231-254, 2012.

Appendix A

PARECER SOBRE PROJETO DE TESE

Documentos: Projeto de tese de Doutorado de autoria de Ana Lúcia Surerus Pitanguy Marques, intitulado "Science and Geography lexicons in English for learners in FUNDAMENTAL 1: A Corpus-Based Investigation", a ser orientado pela Profa. Dra. Deise Prina Dutra junto à Linha de Pesquisa em Ensino/Aprendizagem de Línguas Estrangeiras, na Área de Concentração em Linguística Aplicada.

Mérito: O projeto em tela tem por objetivo realizar uma investigação baseada em instrumentos e aportes metodológicos da Linguística de Corpus para o desenvolvimento de materiais e estratégias didáticas compatíveis com o ensino de língua inglesa para aprendizes em idade do ensino fundamental 1, em contexto da escola regular.

Trata-se, ainda, de uma proposta de estudo que se insere explicitamente no contexto dos debates pedagógicos do ensino integrado de língua e conteúdo (ou *CLIL*, acrônimo do inglês *Content and Language Integrated Learning*). O CLIL é uma perspectiva pedagógica emergente na arena do ensino/aprendizagem de línguas adicionais que contempla uma multiplicidade de abordagens, aplicadas a todos os níveis da educação formal, em torno das quais há grandes expectativas quanto a sua eficácia para que o ensino de línguas adicionais em ambientes escolares/acadêmicos seja capaz de conduzir os aprendizes a níveis robustos de proficiência em língua adicional. Essa perspectiva vem gerando profícua literatura científica, especialmente a partir de estudos de suas implementações em países europeus. Um mérito da presente proposta é tratar-se de um estudo sobre o CLIL em contexto brasileiro, o que pode ser compreendido como uma resposta rápida de nossa comunidade de pesquisa a uma tendência muito recente, mas já notória e muito influente, no ensino de línguas contemporâneo

Subjacente a esse objetivo geral encontra-se a meta especifica de implementação de dispositivos pedagógicos voltados ao desenvolvimento lexical contextualizado e organizado em redes de recorrência co-textual configuradoras de alta idiomaticidade e alta frequência. Assim, o projeto em tela assenta-se com coerência em um prisma lexicalista da organização da linguagem humana, não pautado pela pressuposição de um componente combinatório separado dos repositórios de memória lexical (memória semântica e, muito provavelmente, memória episódica). Trata-se de um ponto de vista que em anos recentes ganhou notória centralidade tantos nos estudos gramaticais quanto nos estudos do processamento da linguagem, sendo a opinião do presente parecerista que um outro aspecto altamente meritório da proposta de estudo é precisamente a translação desse prisma para uma proposta de intervenção educacional.

O projeto coteja os objetivos delineados com uma revisão bibliográfica, de ampla abrangência histórica, tanto dos corolários pedagógicos de abordagens de ensino do tipo *data driven learning* (baseadas em métodos da Linguística de Corpus) quanto de questões da aquisição do léxico em línguas adicionais e seus impactos sobre habilidade específicas dos aprendizes/usuários dessas línguas. O planejamento metodológico prevê um estudo em 5 etapas, que comtemplam desde a composição de um corpus para o estudo quanto o delineamento de atividades didáticas em *data-driven learning*. O cronograma é coerente e prevê a elaboração de uma tese dentro dos prazos usuais.

.

Parecer: Pelo o exposto acima, sou s.m.j. favorável à aprovação sem restrições do projeto de dissertação em tela.

Prof. Dr. Ricardo Augusto de Souza

Belo Horizonte, 9 de setembro de 2021

Appendix B

POSLIN·UFMG

Universidade Federal de Minas Gerais
Faculdade de Letras
Programa de Pós-Graduação em Estudos Linguísticos

D E C L A R A Ç Ã O

Declaro, para os devidos fins, que o projeto de pesquisa da ***doutoranda Ana Lúcia Surerus Pitanguy Marques***, intitulado ***"LÉXICOS DE CIÊNCIA E GEOGRAFIA NO FUNDAMENTAL I: UMA INVESTIGAÇÃO BASEADA EM CORPORA"***, desenvolvido sob a orientação da professora ***Deise Prina Dutra*** foi aprovado em reunião do Colegiado de 13 de dezembro de 2021, e teve como parecerista o professor ***Ricardo Augusto de Souza***.

Belo Horizonte, 16 de dezembro de 2021.

Este documento eletrônico dispensa carimbo e assinatura. Sua autenticidade pode ser comprovada através da ferramenta de verificação de autenticidade de documentos, disponível na página do Programa de Pós-Graduação em Letras - Estudos Linguísticos - FALE - UFMG, neste endereço: http://www.poslin.letras.ufmg.br/confdoc.php

Documento emitido às 17:15 de 16 de dezembro de 2021
Código de verificação de autenticidade: PGPOSLIN916122021171512NBVEQ

Appendix C

Centro Pedagógico UFMG
Escola de Educação Básica e Profissional
Universidade Federal de Minas Gerais
Setor NAPQ/CP/UFMG

Interessados/as: Ana Lúcia Surerus Pitanguy Marques
Parecerista: Dr. Júlio César Virgínio da Costa – Sub-coordenador NAPq/CP/UFMG
Projeto de Pesquisa: "Léxicos de Ciência e Geografia no FUNDAMENTAL 1: uma investigação baseada em *corpora*"
Unidade: Centro Pedagógico

PARECER CONSUBSTANCIADO

HISTÓRICO

O projeto de pesquisa "Léxicos de Ciência e Geografia no FUNDAMENTAL 1: uma investigação baseada em *corpora*", " tem com proponente Ana Lúcia Surerus Pitanguy Marques, estudante do doutorado na Faculdade de Letras da UFMG. A proponente protocolou a solicitação de registro do projeto de pesquisa ao NAPq/CP/UFMG, em 3/3/2022, apresentando os seguintes documentos: formulário de solicitação da pesquisa, resumo da pesquisa, carta de apresentação da orientadora, TALE para pesquisa com menores de idade, TCLE e posteriormente, o projeto de pesquisa.

O estudo tem como objetivo geral: realizar uma investigação baseada em instrumentos e aportes metodológicos da Linguística de Corpus para o desenvolvimento de materiais e estratégias didáticas compatíveis com o ensino de língua inglesa para aprendizes em idade do ensino fundamental 1, em contexto da escola regular.

Os objetivos específicos elencados são: 1. Quais são as palavras de conteúdo relacionadas aos tópicos mais frequentes em Ciências Sociais, Ciências e Geografia nas séries 2ª. – 6ª. da Fundamental I? 2. Quais são as combinações mais frequentes de palavras e colocados (3 – 4 gramas) nessas matérias nas séries 2ª. – 6ª. da Fundamental I? 3. Há variação significativa das palavras de conteúdo ao longo das séries iniciais do Fundamental I? 4. Como as atividades baseadas em corpora e nos princípios da aprendizagem movida por dados (JOHNS, 1991) aplicadas em sala de aula contribuíram para a aquisição de vocabulário?

Trata-se de uma pesquisa na qual as atividades que serão criadas terão como premissa a sua execução dentro da abordagem DDL - *data-driven learning* - ou ensino por meio de dados de *corpora*. O que difere essa abordagem de uma abordagem comumente utilizada pela comunidade escolar é que ela trabalha com a língua, gramática ou vocabulário neste caso, através da exposição dos alunos aos dados obtidos de textos autênticos por um *software*. É uma abordagem indutiva onde espera-se que os alunos observem o comportamento das palavras ou frases em seus contextos

Centro Pedagógico UFMG
Escola de Educação Básica e Profissional
Universidade Federal de Minas Gerais
Setor NAPQ/CP/UFMG

originais – linhas de concordância - e, após reiterado uso em atividades variadas e interativas, possam inseri-las em sua produção escrita.

O vocabulário a ser trabalhado será obtido de um *corpus* topicalizado e as atividades escolhidas poderão ser com 'nuvem de palavras', *flashcards* (fichas ou cartões), trabalho em pares ou grupos para maior negociação de significado e retenção da forma, imagens quando necessário e também excertos dos textos originais. Após a apresentação do vocabulário-chave e recorrência da sua exposição nas atividades, será pedido aos alunos que produzam sentenças descrevendo imagens que deverão conduzir os alunos à utilização do vocabulário estudado. Um pequeno teste para a avaliação da retenção será feito após uma semana de aulas.

MÉRITO

Trata-se de uma pesquisa que tem como campo investigativo a prática de ensino de Língua Inglesa na Educação Básica, especificamente, no ensino fundamental.

O projeto de pesquisa é relevante e pertinente para a produção acadêmica na área escolar e da área do ensino de línguas estrangeiras. Sugerimos clarear os objetivos específicos para uma melhor compreensão dos passos da pesquisa.

VOTO

O projeto destaca-se pela relevância do tema e a contribuição no campo do estudos das línguas estrangeiras e para a escola. Nesse sentido, sou, S.M.J. **favorável à aprovação da proposta do projeto analisado.**

Belo Horizonte, 28 de março de 2022.

Julio Cesar Virginio da Costa:91767288620
Assinado de forma digital por Julio Cesar Virginio da Costa:91767288620
Dados: 2022.03.28 09:52:24 -03'00'

Prof. Dr. Júlio César Virginio da Costa
Subcoordenador do NAPQ/CP/UFMG

APROVADO *AD REFERENDUM* EM 28/03/2022.

Marcos Elias Sala
Assinado de forma digital por Marcos Elias Sala
Dados: 2022.03.28 15:58:19 -03'00'

Appendix D **Pre- and posttests – *Neighborhood* (T1)**

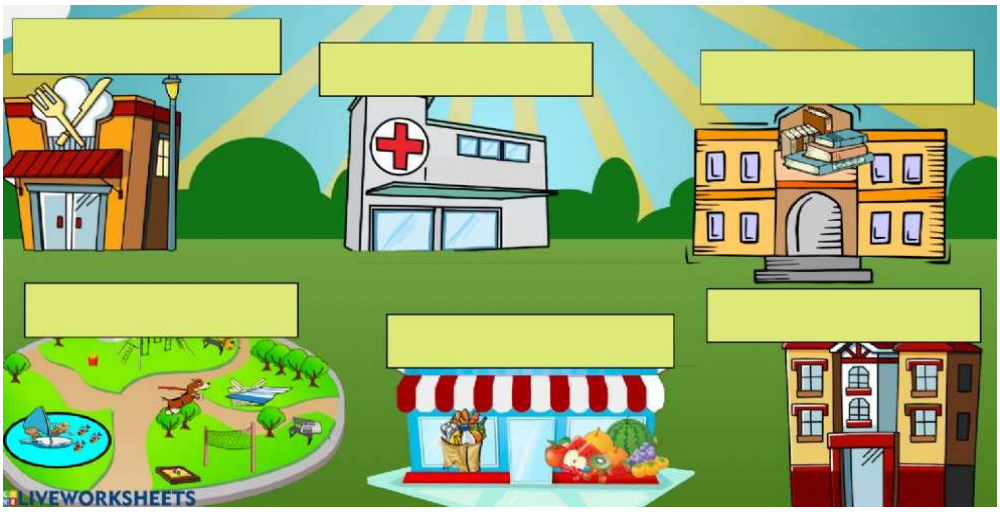
UNIVERSIDADE FEDERAL DE MINAS GERAIS
Escola de Educação Básica e Profissional
Centro Pedagógico - Núcleo de Línguas Estrangeiras

Student _____ Age _____ Grade _____

Pre-test - *Neighborhood* - April 2022

Elementary School | House | Gas Station | Hospital | Bakery
Movie Theater | Bank | Park | Shop | Restaurant | Drugstore
Supermarket | Garden Center | Office | Fire Station | Pet shop
Library | High School | Mall | Police Station | Post Office

1- Look at the words and label the buildings:



2- Choose 3 buildings and write sentences about them:

a- _____

b- _____

c- _____

3- Match the two columns:

(1) Hospital	( ) fireman
(2) School	( ) police officer
(3) Police station	( ) doctors and nurses
(4) Post office	( ) principal
(5) Bakery	( ) postman
(6) Fire station	( ) baker

4- Look at the picture about this town. Which buildings do you know? Write the names you can remember:

_____ _____ _____

_____ _____ _____

_____ _____ _____

5- Complete the sentences to describe your neighborhood:

I live near the _____ and _____

There are _____

There is _____

UNIVERSIDADE FEDERAL DE MINAS GERAIS

Escola de Educação Básica e Profissional

Centro Pedagógico - Núcleo de Línguas Estrangeiras

Student _____ Age _____ Grade _____

Post-test – *Neighborhood* – April 2022

Elementary School | House | Movie Theater | Bank | Park | Supermarket | Garden Center | Library | High School | Mall

Gas Station | Hospital | Bakery | Shop | Restaurant | Drugstore | Office | Fire Station | Pet shop | Police Station | Post Office

1- Look at the words and label the buildings:

2- Read the concordance lines below.
Then, choose **3 buildings (exercise 1) and write sentences <u>about them</u>:**

- There are many different types of map. This is because we use different maps for different reasons.
- find the best road to your friend's house
- find your way around a nature park
- City block with homes and stores
- Do you have friends in your neighborhood?
- Tall apartment buildings where many people live
- They may have museums, libraries, and parks.
- People live, work, learn, and have fun close to one another in cities
- They may be able to walk to school, the post office, the library, and stores.
- They may also use public transportation to get to different parts of the city.
- Tall apartment buildings where many people live
- There is a library nearby
- When you have a picnic at the park, you clean up after yourself
- A park map, for example, help you plan

a- _____

b- _____

c- _____

3- Match the two columns:

(1) Hospital	( ) fireman
(2) School	( ) police officer
(3) Police station	( ) doctors and nurses
(4) Post office	( ) principal
(5) Bakery	( ) postman
(6) Fire station	( ) baker

4- Look at the picture about this town. Which buildings do you know? Write the names you can remember:

_____ _____ _____

_____ _____ _____

_____ _____ _____

5- Complete the sentences to describe your neighborhood:

I live near the _____ and _____

There are _____

There is _____

Appendix E **Pre- and posttests – *Animals* (T2)**

UNIVERSIDADE FEDERAL DE MINAS GERAIS

Escola de Educação Básica e Profissional

Centro Pedagógico - Núcleo de Línguas Estrangeiras

Student _____Age: _____ Grade: _____

Pre-test – *Animals* – May 2022

1. Look at the pictures and label them with the words:

octopus **giraffe** **shark** **parrot** **rabbit** **snake** **lion** **chimpanzee** **horse**

_____ _____ _____

_____ _____ _____

_____ _____ _____

2- Choose **3 pictures and write sentences** **<u>with the words you selected:</u>**

Example: The **shark** lives in the ocean. The **shark** is white and black. **Sharks** eat fish.

a- _____

b- _____

c- _____

3- Complete **<u>each group of sentences</u>** **<u>with the same word</u>** from the list or with the ones you remember:

a) The _____ is a big cat.

The _____ lives in the wild in Africa and India.

The _____ hunts during the day.

The _____ have a beautiful mane around their head.

b) The _____ live in the water, in rivers and in a bowl in the houses.

Some _____ are gray, others are yellow, blue or red.

The _____ which live in the Amazon have a dangerous bite.

c) The _____ live on farms and on the plains.

Some children like to ride the domestic _____.

The _____ can be brown, black, gray and also beige.

4- Make or complete the sentences about the **octopus**

_____ **(6 arms and 2 legs)**

_____ **(mollusk)**

The octopus _____ in the water in the _____

5- Let's classify the animals. Read the names and put them in the right columns below:

- tiger, dog, lion, fish, cat, cow, elephant, horse, snake, shark, frog, panda bear, giraffe, parrot, rabbit

Domestic	**Pets**	**Wild**

UNIVERSIDADE FEDERAL DE MINAS GERAIS

Escola de Educação Básica e Profissional

Centro Pedagógico - Núcleo de Línguas Estrangeiras

Student _____Age: _____ Grade: _____

Posttest – *Animals* – May 2022

2. Look at the pictures and label them with the words:

octopus giraffe shark parrot rabbit snake lion chimpanzee horse

_____

_____

_____

_____

_____

_____

_____

_____

_____

2- Read the examples and choose **3 pictures and write sentences** with the words you selected:

- Mammals have fur or hair and feed on milk. Humans are mammals too.
- Butterflies, ants and bees are all insects. Insects have 6 legs.
- Crickets are jumping insects.
- Mosquitoes are insects that are found almost everywhere in the world.
- Alligators and crocodiles are powerful animals with powerful tails.
- Fish are a kind of animal that lives in the water.
- Birds are animals that have feathers.
- Many people have birds as pets.
- Lions live in parts of Africa and India.
- The tiger is the largest of the cats. They are very strong and good hunters.
- Some birds cannot fly at all.
- Penguins are birds that cannot fly.
- Frogs are small animals that can jump very well.
- Zebras are animals that have strong legs to run.

Example: The **shark** lives in the ocean. The **shark** is white and black. **Sharks** eat fish.

a- _____

b- _____

c- _____

3- Complete **each group of sentences** **with the same word** from the list or with the ones you remember:

a) The _____ is a big cat.

The _____ lives in the wild in Africa and India.

The _____ hunts during the day.

The _____ have a beautiful mane around their head.

b) The _____ live in the water, in rivers and in a bowl in the houses.

Some _____ are gray, others are yellow, blue or red.

The _____ which live in the Amazon have a dangerous bite.

c) The _____ live on farms and on the plains.

Some children like to ride the domestic _____.

The _____ can be brown, black, gray and also beige.

4- Make or complete the sentences about the **octopus**

_____ **(6 arms and 2 legs)**

_____ **(mollusk)**

The octopus _____in the water in the _____

5- Let's classify the animals. Read the names and put them in the right columns below:

- Tiger, dog, lion, fish, cat, cow, elephant, horse, snake, shark, frog, panda bear, giraffe, parrot, rabbit

Domestic	**Pets**	**Wild**

Appendix F **Homework – *Neighborhood* (T1)**

UNIVERSIDADE FEDERAL DE MINAS GERAIS

Escola de Educação Básica e Profissional

Centro Pedagógico - Núcleo de Línguas Estrangeiras

Name:_____ Age:____ Grade:_____

PLACES IN A CITY

1. Look at the map. Then, answer the questions.

1. Is there a supermarket? _____

2. Are there three cars? _____

3. Is there a swimming pool? _____

4. Is there a dog? _____

5. Are there two parks? _____

6. Are there two shops? _____

7. Is there a school bus? _____

8. Is there a hotel? _____

9. Are there two supermarkets? _____

10. Is there a zoo? _____

Remember!

Is there...?

- ✓ Yes, there is.
- ✗ No, there isn't.

Are there...?

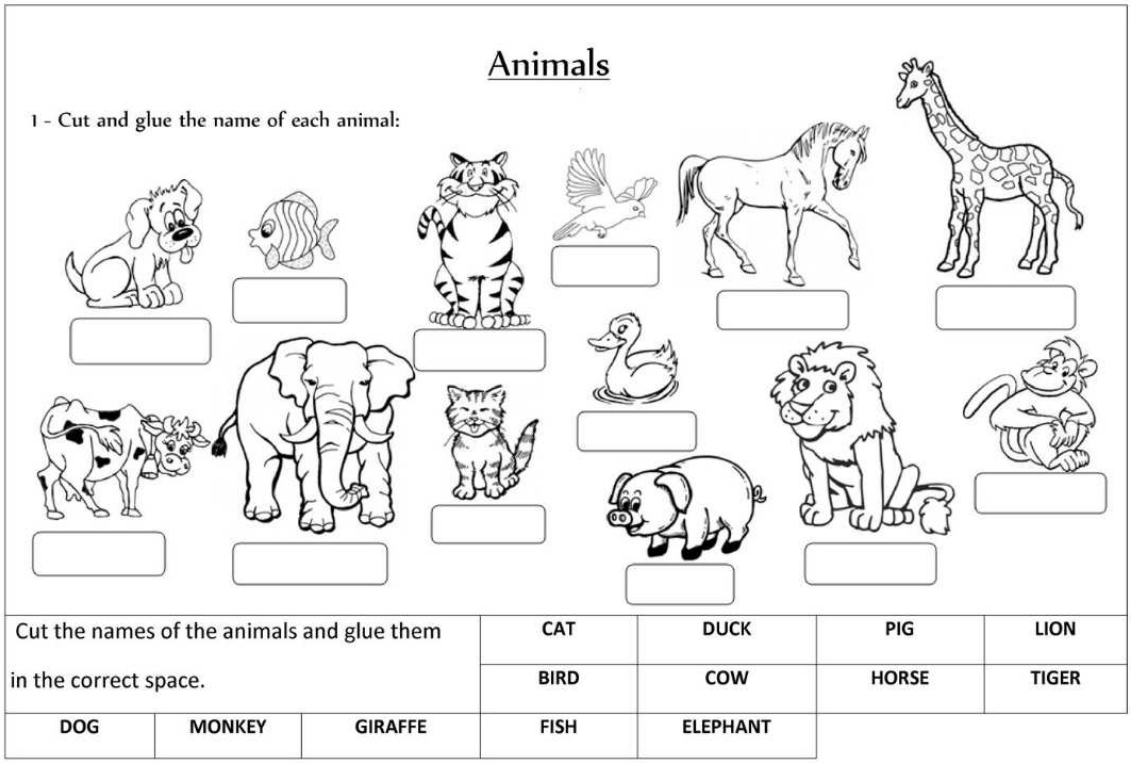
- ✓ Yes, there are.
- ✗ No, there aren't

Appendix G **Homework – *Animals* (T2)**

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Escola de Educação Básica e Profissional
Centro Pedagógico - Núcleo de Línguas Estrangeiras

Name: _____ Age:____ Grade:_____

1- Label the animals and write sentences about **3 you like most:**



1. Which animals do you know? (at least 3 - 0.5 each)

2. Which ones can you find **in the zoo** or **at home**? (maximum 8)

In the Zoo	At home pets – domestic

3. Name and describe 5 of them. Say how they move (swim, fly, run or walk) (0.5 each)

_____

_____

_____

Appendix H – Video transcription of *Our Neighborhood* (Topic 1)

Hello friends, how are you

in this video we will learn about our **neighborhood**

neighborhood means places near us

the area around our house is called our neighborhood

the houses built close to each other make up our neighbors

our neighborhood has many services now we will go through these services

this is the **supermarket**

it has many shops we go to the **market** to buy things

this is a **hospital**

it has many doctors and nurses

we came here when we fall ill

this is my **school**

there are many teachers in a school to teach us

the principal is the head of a school

this is a **police station** the police help in keeping law and order in our area

police catches criminals

this is a petrol pump

we fill fuels in our vehicles at petrol pump

gas, petrol, diesel all these are available at pump

this is a **post office**

we get postcards letters and envelopes here

a postman delivers letters

this is a bank

people keep money and valuable here

people deposit and withdraw money from **bank**

this is a **bakery**

baker makes tasty biscuits and cookies for us

cakes, pastries are specialties of a bakery

this is a fruits and vegetables vendor

we get fresh fruits and vegetables here

vendor brings fruits and vegetables from farmers

this is a fire station

there are fire engines in the fire station

firemen helped to put out fires

this is a library

library contains different types of books

we came here to read books

this is a park

children came here to play

people came here for walks

this is a pharmacy

pharmacy

chemist gives medicines as per prescription

this is a bus stop

we use bus to travel

different routes buses are available at bus stand

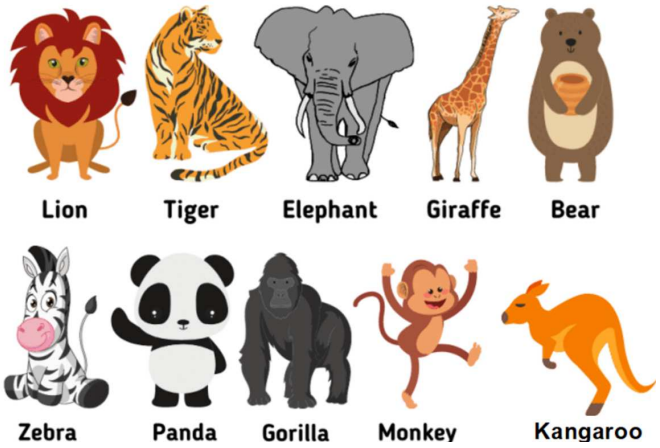
this is a theater

we see movies and drama in theater

live stage shows are also part of theaters.

(Available at: https://www.youtube.com/watch?v=OQxRVOG10ZA&t=9s. Accessed on: Feb 1st 2022.)

Appendix I **Bingo cards samples – *Animals* (Topic 2)**[66]



Instructions to the teacher:

Esses 'cards' serão ótimos para que trios de alunos recebam um 'card' de animais + nomes e possam receber cartões individuais de bingo e brincar entre eles para ver quem tem o 'card de bingo' com mais nomes relacionados ao 'card' de animais + nomes.

Pode haver drilling inicialmente dos nomes e em associação às imagens, para depois o jogo de bingo.

Para cada acerto o alunos precisa ler o nome corretamente e checar o animal no card. (Drilling + noticing)

[66] It is possible to resort to different sites using artificial information (AI) which supply images for any text submitted. Examples: Midjourney, Bing Image Creator, Craiyon, NightCafe, Dream by Wombo, Adobe Firefly.

Appendix J **More detailed lesson plan (Topic 1)**

Title: Getting to know your neighborhood / town

Level of students: Year 4, 5 and 6 – Elementary School

Aims: To work with vocabulary related to directions, places of interest in the neighborhood and reinforce the awareness of correct spelling of most frequent words in lessons: NOUNS, VERBS, PLACES AND DIRECTIONS.

Class time: 60'

Preparation time: A couple of hours to get the context-related pictures, to select the coursebook Units to build the corpus, to prepare and print the concordance lines for the Word Wall, to select and create the word clouds and the word cards, to add other proper nouns seen elsewhere and related to the topic.

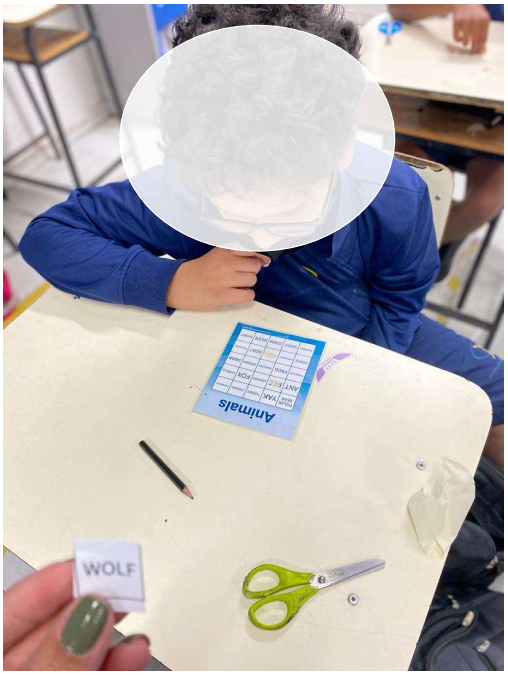
Resources/Materials: See in the Plan for each stage's material.

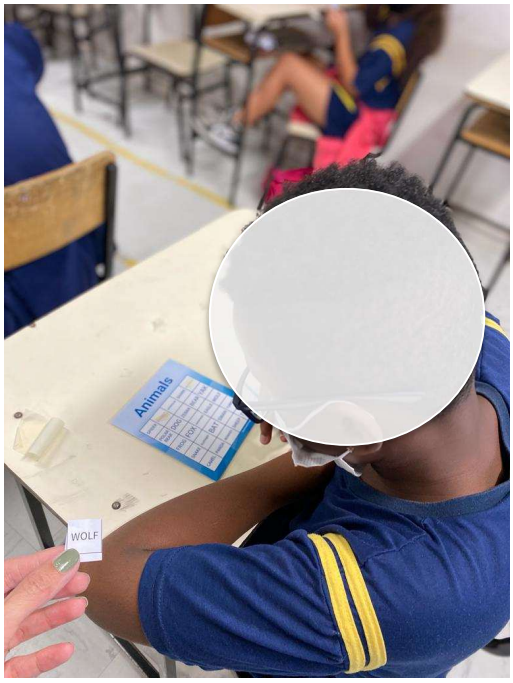
Stages / steps	Procedures	Time	Patterns of Interaction	Materials
Warm – up	-T shows pictures of different neighborhoods by different authors to spark interest and trigger learners' motivation to draw their own neighborhood later in the lesson. -These pictures are also the context for the vocabulary they will be working with. - T asks simple questions and / or describes those drawings using the focus vocabulary for places and directions.	10'	Teacher- Learners (whole- class)	Pictures and drawings are shown

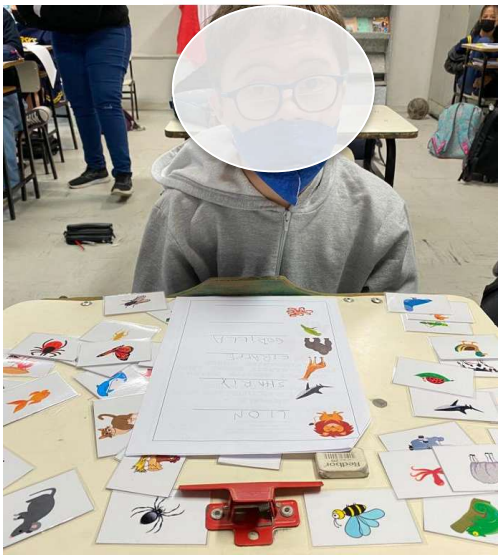
Vocab I **(game)**	-T hands out 7 copies of Word clouds with some of the most frequent words (KWICs) from the 5 Units of the coursebook: verbs, directions and places. - Competition: learners need to find as many words as possible in 5 minutes, write them down and count. The trio with most words wins. - Feedback to class. - Learners say the words out loud and pronunciation is checked	10’	Trios T - Ss S - T	Printouts of Word clouds
Vocab II	- Teacher lays the white construction paper on the floor. - Each learner receives some concordance lines (printouts) selected by the T with one of the most frequent KWICS in context and any additional word related to the Brazilian reality but in English. - Learners read the sentences, then say the KWIC aloud and nominate another student to pick up the corresponding KWIC in the black cards and place it on the Word Wall (the white board). - Feedback - Teacher checks understanding and hangs the poster on the wall to be used as a reminder of vocab in the next stage.	20’	Individual Open pairs	Construc tion paper Printouts Cards with KWICs

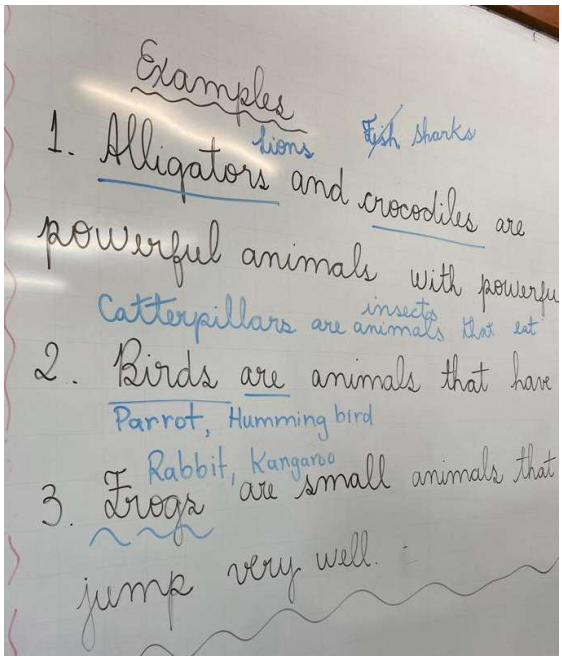
Free Practice: (Speaking practice giving locations)	- Learners receive the drawing of a town center. Half of the class gets A and the other half gets B to create the interest in working in pairs. - They complete the maps choosing words from the Word Wall and work with a partner (A + B) asking and answering questions about their drawings. They should use directions when necessary checking the Compass rose. - Sample questions on the board: - What is this building? - Where is your house? - Is it south of the school? And the supermarket? - Is it near or far from your house? Is it in the West?	10' 10'	Individual Pair work	Drawings A and B Word Wall
HWork **Personalization** (Writing practice)	**- Learners take their drawings home and write 5 sentences using the KWICS to describe places and directions.**		Individual	

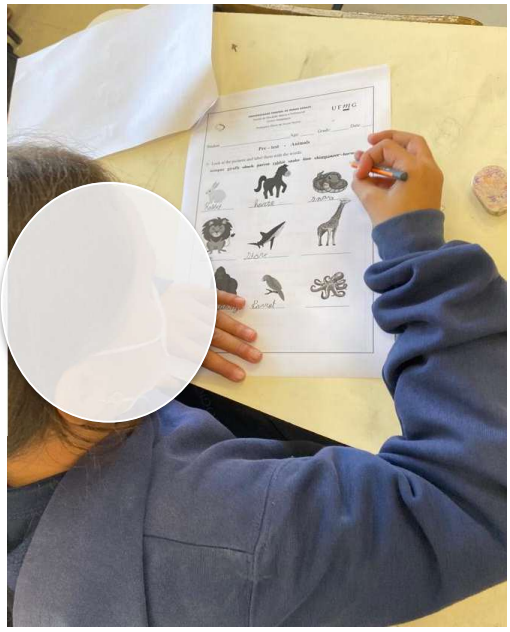
Appendix K **Learners in action in the classroom: photographs**











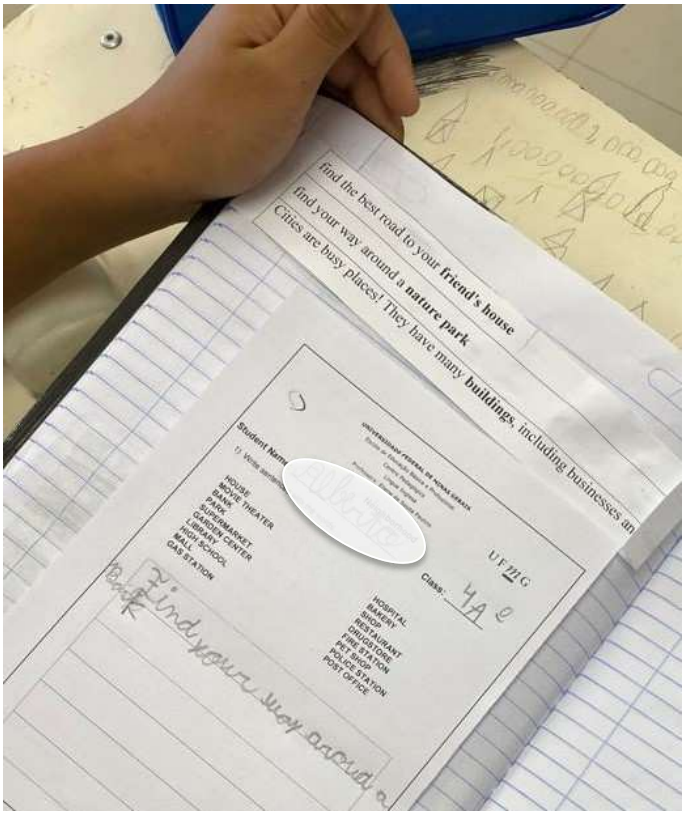

OWL


MY TOWN
shopping mall
park
bank
elementary school
shop office
garden center
pet store
movie theater
high school
restaurant
supermarket
fire station
hospital
drugstore
bakery
library
gas station
post office
myfreebingocards.com


There is a library nearby.
PEOPLE LIVE IN A SINGLE HO
The house you live and your school
pet store
library
post office
park
bakery
shop office
hospital
fire station
garden center
movie theater


The part of a city or town that is close to your home is called your neighborhood.
The stores you use and your school are likely to be in your neighborhood.
They may have museums,libraries, and parks.


Planner Rental
There is a library nearby
PEOPLE LIVE IN A SINGLE HOUSE WITH A YARD, OR IN AN APARTMENT.


find the best road to your friend's house
find your way around a nature park
Cities are busy places! They have many buildings, including businesses an
Student Name
HOUSE
MOVIE THEATER
BANK
PARK
SUPERMARKET
GROCERY CENTER
LIBRARY
MALL/SCHOOL
GAS STATION
HOSPITAL
BAKERY
SHOP
RESTAURANT
DRUGSTORE
FIRE STATION
POLICE STATION
POST OFFICE
Class: 4A
Back
Find your way around

Appendix L Most frequent words: COREL-GEO and 3- and 4-grams

1- Most frequent words

Search

▼ Corpus COREL_GEO ▼ Frequency ▼ Dispersion ▼ Ty

Type	▼ Frequency: 01 - Freq	Dispersion: 01_CV
the	6375.000000	0.447758
of	2569.000000	0.650346
and	2453.000000	0.677885
is	2046.000000	0.747554
a	2034.000000	0.871552
to	1859.000000	0.800018
in	1710.000000	0.752914
are	1427.000000	0.957705
it	887.000000	1.131860
that	800.000000	1.051627
you	669.000000	1.799714
on	592.000000	1.396621
they	589.000000	1.549366
can	549.000000	1.629760
water	514.000000	2.026305
for	505.000000	1.509535
people	491.000000	1.674919
from	470.000000	1.451002
this	467.000000	1.509161
or	437.000000	1.771113
there	431.000000	1.555347
as	422.000000	1.692086
ocean	416.000000	3.107675
many	383.000000	1.572766
have	372.000000	1.995058
animals	364.000000	2.484057
what	355.000000	1.738440
earth	349.000000	2.362486
be	342.000000	1.984132
map	328.000000	2.051140
an	315.000000	2.033049
live	315.000000	1.784771
at	301.000000	1.601835
by	286.000000	1.966256
some	281.000000	1.914100
land	275.000000	2.124473
these	268.000000	2.007051
their	265.000000	2.241445
with	262.000000	1.933580
which	257.000000	2.582361
all	252.000000	2.005086
world	250.000000	3.308399
then	248.000000	1.442506
plants	246.000000	2.657944

Search

▼ Corpus COREL_GEO ▼ Frequency ▼ Dispersion ▼ Ty

Type	▼ Frequency: 01 - Freq	Dispersion: 01_CV
called	244.000000	2.111608
we	242.000000	3.216093
but	238.000000	2.010886
also	236.000000	1.685830
how	228.000000	1.850643
has	225.000000	2.156396
about	223.000000	2.258331
like	214.000000	2.590166
south	213.000000	2.610037
north	213.000000	2.598919
answer	213.000000	1.572286
river	209.000000	3.304423
where	206.000000	2.263816
when	205.000000	2.218364
different	204.000000	2.186691
one	204.000000	2.092779
study	203.000000	1.947578
do	203.000000	2.215198
other	203.000000	2.086241
not	196.000000	1.999193
very	196.000000	2.662030
up	182.000000	2.447462
more	182.000000	2.371151
america	181.000000	3.437988
your	181.000000	3.434351
most	178.000000	2.695834
each	175.000000	2.289210
use	173.000000	2.213553
into	172.000000	2.554736
questions	167.000000	1.619144
its	165.000000	2.498986
them	164.000000	2.705288
around	164.000000	2.545620
places	160.000000	2.555130
found	160.000000	2.736643
natural	160.000000	3.316243
so	159.000000	2.766422
read	159.000000	1.737489
help	158.000000	2.881209
trees	154.000000	3.424070
large	151.000000	2.890140
things	150.000000	2.866926
place	149.000000	2.976124
below	146.000000	1.781074
than	141.000000	2.609264
look	140.000000	2.177096
planet	139.000000	4.822710
city	135.000000	3.637464
mountains	133.000000	3.962335
oceans	132.000000	4.556961
states	131.000000	3.347345

Search

▼ Corpus COREL_GEO ▼ Frequency ▼ Dispersion ▼ Ty

Type	▼ Frequency: 01 - Freq	Dispersion: 01_CV
photo	131.000000	2.106618
why	129.000000	2.620037
text	128.000000	1.867053
over	128.000000	3.016944
make	126.000000	2.670381
because	125.000000	3.099753
such	123.000000	2.802964
desert	122.000000	4.857702
our	121.000000	3.933077
rivers	120.000000	3.673305
types	120.000000	2.986169
was	119.000000	3.881738
sun	118.000000	5.223588
rain	114.000000	4.865803
part	114.000000	3.521843
will	114.000000	3.004461
find	112.000000	2.799642
food	111.000000	4.835477
africa	111.000000	4.544387
if	109.000000	3.253099
forest	109.000000	4.606801
two	108.000000	2.903655
areas	107.000000	3.539433
were	106.000000	3.911244
need	105.000000	3.152950
only	105.000000	3.371087
may	105.000000	3.577469
forests	105.000000	4.560624
out	104.000000	3.571507
made	103.000000	3.099626
community	103.000000	5.629545
circle	103.000000	2.782508
soil	103.000000	4.673530
way	102.000000	3.023230
new	102.000000	4.176368
continent	100.000000	3.705299
time	99.000000	3.321234
area	98.000000	2.988245
rock	98.000000	6.177010
pacific	97.000000	4.151692
life	97.000000	3.754771
home	97.000000	3.339429
largest	96.000000	3.144889
mountain	96.000000	3.888821
much	95.000000	3.113137
it's	95.000000	4.949404
see	94.000000	3.751967
living	94.000000	3.920410
cold	93.000000	3.909800
asia	92.000000	4.110210
hot	91.000000	3.815578

2 Most frequent 3-grams

Search

▼ Corpus COREL_GEO ▼ Frequency ▼ Dispersion ▼ Ty

Type	▼ Frequency: 01 - Freq	Dispers
answer the questions	162.000000	1.673367
then answer the	123.000000	1.973365
read the text	120.000000	1.886886
study the photo	106.000000	2.022513
the photo then	102.000000	2.051204
photo then answer	88.000000	2.175096
plants and animals	78.000000	3.481472
the text and	77.000000	2.297180
text and study	73.000000	2.371971
and study the	73.000000	2.371971
the united states	62.000000	4.245533
part of the	62.000000	4.748069
look at the	61.000000	2.806548
of the world	49.000000	4.317296
live in the	48.000000	4.165676
in the world	47.000000	4.136238
from the sun	44.000000	5.948265
on the map	44.000000	3.654467
a lot of	42.000000	4.723591
the natural world	42.000000	5.559559
there are many	41.000000	5.017760
one of the	40.000000	5.229525
of the earth	35.000000	5.804702
animals that live	34.000000	4.677408
that live in	34.000000	5.732781
the human world	33.000000	6.480326
the questions the	32.000000	3.795000
different types of	32.000000	5.507190
circle the answer	32.000000	5.243624
some of the	31.000000	6.242278
the atlantic ocean	31.000000	6.729542
the answer to	31.000000	5.237746
is the largest	30.000000	4.685818
answer to each	30.000000	5.254863
can be found	30.000000	6.057534
study the map	30.000000	4.643025
the pacific ocean	29.000000	8.933757
text study the	29.000000	4.192359
what is the	29.000000	5.699734
bodies of water	29.000000	5.568719
the text study	29.000000	4.192359
to each question	29.000000	5.310329
at the map	28.000000	4.230520

Search

▼ Corpus COREL_GEO ▼ Frequency ▼ Dispersion ▼ Ty

Type	▼ Frequency: 01 - Freq	Dispers
is called a	28.000000	6.478677
found in the	28.000000	5.634385
the amazon rainforest	28.000000	10.105105
body of water	27.000000	5.174610
it is a	25.000000	4.759432
do you think	24.000000	5.754694
different kinds of	24.000000	6.377364
a map of	24.000000	5.782442
there is a	24.000000	5.005948
made up of	24.000000	6.104736
the earth is	24.000000	8.257587
what is a	23.000000	4.870037
the name of	23.000000	6.150523
most of the	23.000000	5.721576
in the ocean	23.000000	6.899400
is part of	22.000000	8.756900
in the united	22.000000	5.826448
is a large	22.000000	5.471376
is one of	21.000000	6.077005
be found in	21.000000	7.671668
it is the	21.000000	6.075649
the map of	21.000000	4.799366
much of the	20.000000	6.641180
directions read the	20.000000	5.043159
is home to	20.000000	5.700239
animals and plants	20.000000	9.364529
look at this	20.000000	5.437830
in the photo	19.000000	5.569569
map of a	19.000000	5.279871
is called the	19.000000	7.263712
are found in	19.000000	7.061376
the plants and	19.000000	6.233045
the compass rose	19.000000	8.459438
in north america	19.000000	8.984202
in the desert	19.000000	8.406160
map of the	19.000000	5.203343
the solar system	19.000000	9.339678
the greenhouse effect	18.000000	11.127980
the surface of	18.000000	7.915276
gulf of mexico	18.000000	8.307563
the earth's surface	18.000000	8.001786
is made up	18.000000	7.170250
to live in	17.000000	7.271216
parts of the	17.000000	6.574826
next to the	17.000000	6.613917
of land that	17.000000	6.802716
of plants and	17.000000	6.775539
this map of	17.000000	5.368876
it is also	17.000000	6.516627
and answer the	17.000000	6.104699
of the natural	17.000000	6.777080

Search

▼ Corpus COREL_GEO ▼ Frequency ▼ Dispersion ▼ Ty

Type	▼ Frequency: 01 - Freq	Dispers
for people to	17.000000	6.239623
there are also	17.000000	5.955269
ocean pacific ocean	16.000000	6.121099
in south america	16.000000	8.020603
to help you	16.000000	5.820930
the top of	16.000000	6.339038
of the earth's	16.000000	7.551724
surface of the	16.000000	8.881093
be able to	16.000000	8.031061
all of the	16.000000	7.966267
the sun and	16.000000	8.841786
around the world	16.000000	8.295201
in the amazon	16.000000	8.713498
do you know	16.000000	7.116797
the map and	16.000000	6.145455
as well as	16.000000	8.414238
surrounded by land	16.000000	6.948143
areas of land	15.000000	6.467360
area of land	15.000000	6.673765
the sun the	15.000000	9.621988
places on earth	15.000000	7.618848
planet from the	15.000000	13.401017
the number of	15.000000	6.625448
the names of	15.000000	6.658830
of the ocean	15.000000	10.705958
is in the	15.000000	8.156782
name of the	15.000000	7.094027
the map below	15.000000	5.866191
out of the	14.000000	8.629915
in the box	14.000000	6.436056
this is a	14.000000	8.429639
it can be	14.000000	7.276471
the questions below	14.000000	6.248786
the types of	14.000000	7.039111
can be used	14.000000	7.380121
how do you	14.000000	6.643544
the things that	14.000000	7.416556
top of the	14.000000	7.909344
an urban community	14.000000	12.019180
are part of	14.000000	8.414596
the ocean is	14.000000	9.177286
map circle the	14.000000	6.995172
of the land	14.000000	7.424789
you want to	14.000000	7.370741
the north pole	14.000000	7.389634
in the tundra	14.000000	10.553145
north and south	14.000000	8.673054
then color the	13.000000	7.623148
you can find	13.000000	8.053778
known as the	13.000000	7.650395
also known as	13.000000	7.306033

3 Most frequent 4-grams

Search

▼ Corpus COREL_GEO ▼ Frequency ▼ Dispersion ▼ Ty

Type	▼ Frequency: 01 - Freq	Dispers
the animals that live	8.000000	9.771125
part of the world	8.000000	9.919834
look at the pictures	8.000000	7.756199
a map of a	8.000000	8.880807
it is the largest	7.000000	9.750459
follow the instructions below	7.000000	8.388738
north america south america	7.000000	12.678117
a biome is a	7.000000	12.888925
the united states and	7.000000	10.312312
map and answer the	7.000000	9.489176
the united states is	7.000000	8.784333
adult to help you	7.000000	9.414728
in the amazon rainforest	7.000000	12.906638
in the middle of	7.000000	8.771191
is a group of	7.000000	11.089686
it is home to	7.000000	9.244732
and non-living things interact	7.000000	21.377558
map then answer the	7.000000	9.630580
a community is a	7.000000	11.466587
heat from the sun	7.000000	14.216162
is the name of	7.000000	9.541251
next to the things	7.000000	9.660842
the map then answer	7.000000	9.630580
the sun and the	7.000000	11.261841
area of land that	7.000000	10.077517
of the earth is	7.000000	14.294742
large areas of land	7.000000	9.872072
to live in the	7.000000	9.672310
do you think the	7.000000	9.809975
an adult to help	7.000000	9.414728
found in north america	7.000000	15.202719
map of the world	7.000000	8.280169
what direction will you	7.000000	11.124628
for a long time	7.000000	9.127110
the compass rose to	7.000000	8.318384
why do you think	7.000000	9.347625
north and south america	7.000000	11.455841
south east and west	7.000000	9.154515
in the solar system	6.000000	19.763812
africa south america australia	6.000000	8.978618
ocean the pacific ocean	6.000000	13.592341
of land on earth	6.000000	9.287486
pacific ocean indian ocean	6.000000	8.954361

Search

▼ Corpus COREL_GEO ▼ Frequency ▼ Dispersion ▼ Ty

Type	▼ Frequency: 01 - Freq	Dispers
a rural community has	6.000000	14.580125
the natural world and	6.000000	8.980047
what do you think	6.000000	11.367328
the hottest year on	6.000000	21.377558
different kinds of plants	6.000000	13.353704
is a place where	6.000000	12.920155
the map and answer	6.000000	10.572925
the map of the	6.000000	9.036906
hottest year on record	6.000000	21.377558
that can be found	6.000000	13.141793
can be used to	6.000000	11.025218
use the compass rose	6.000000	9.010327
direction will you go	6.000000	12.332731
why or why not	6.000000	12.474493
from the sun and	6.000000	12.619601
plants and animals in	6.000000	9.594744
north america europe asia	6.000000	9.849050
an animal that eats	6.000000	15.099669
from the sun the	6.000000	16.104876
from the human world	6.000000	11.493040
for people to live	6.000000	10.111738
a place where people	6.000000	12.920155
what is the name	6.000000	9.802253
then follow the instructions	6.000000	9.022133
the sun does not	6.000000	13.947941
will never run out	6.000000	9.770049
out of the ground	6.000000	11.062587
north south east and	6.000000	9.874768
you are at the	6.000000	11.170426
a large body of	6.000000	10.342798
this map of the	6.000000	9.264084
study the map then	6.000000	10.510626
can be found on	6.000000	11.949138
water surrounded by land	6.000000	11.535349
light from the sun	6.000000	11.403224
you want to go	6.000000	11.388619
to go to the	6.000000	12.592216
the north and south	6.000000	11.803271
in the form of	6.000000	11.433027
third planet from the	6.000000	15.370238
of water surrounded by	6.000000	11.535349
is the third planet	6.000000	15.370238
is an area of	6.000000	10.128249
the pacific ocean is	6.000000	18.644050
the third planet from	6.000000	15.370238
the middle of the	6.000000	10.170986
at the pictures below	6.000000	8.750978
a globe is a	5.000000	10.501049
biome is a large	5.000000	10.977189
one place to another	5.000000	10.806128
of the most common	5.000000	10.993637

Search

▼ Corpus COREL_GEO ▼ Frequency ▼ Dispersion ▼ Ty

Type	▼ Frequency: 01 - Freq	Dispers
how can you tell	5.000000	11.903515
the number of people	5.000000	10.800574
also known as the	5.000000	11.540360
from the word box	5.000000	10.388749
types of plants and	5.000000	11.023186
at the top of	5.000000	10.838062
the human world and	5.000000	13.921762
in danger of extinction	5.000000	21.377558
is shaped like a	5.000000	9.851092
there are more than	5.000000	12.350328
the climate of the	5.000000	11.418784
different kinds of animals	5.000000	12.594485
water is called a	5.000000	14.634259
a lot of water	5.000000	14.179932
fossil fuels like coal	5.000000	10.828269
sun and the moon	5.000000	13.092477
a river is a	5.000000	11.194286
the questions the great	5.000000	9.677537
the human world the	5.000000	9.990866
energy from the sun	5.000000	12.932429
is a body of	5.000000	12.499975
at a map of	5.000000	11.369975
be used to make	5.000000	14.990869
are two types of	5.000000	17.193542
places on earth are	5.000000	13.549869
body of water that	5.000000	11.369823
read about it the	5.000000	11.040262
animals that can be	5.000000	14.200724
wouldn't be able to	5.000000	10.369338
in the box below	5.000000	9.660507
place where people live	5.000000	12.425597
europe north america asia	5.000000	10.045009
very large areas of	5.000000	10.211040
interact with each other	5.000000	21.377558
ocean pacific ocean indian	5.000000	9.860174
would you like to	5.000000	11.148388
rose to help you	5.000000	9.679941
seven very large areas	5.000000	10.211040
pacific ocean atlantic ocean	5.000000	10.608833
compass rose to help	5.000000	9.679941
did you know that	5.000000	14.375156
things interact with each	5.000000	21.377558
from one place to	5.000000	10.806128
continents there are seven	5.000000	10.862150
areas of land on	5.000000	10.211040
want to go to	5.000000	12.411903
is a renewable resource	5.000000	11.643612
and the human world	5.000000	11.139154
there are seven very	5.000000	10.211040
the pacific ocean and	5.000000	15.086848
that are part of	5.000000	13.606437

Source: #LancsBox 6.0.

Appendix M Most frequent words: COREL-SCI and 3- and 4-grams

1 Most frequent words

▼ Corpus COREL_SCI ▼ Frequency ▼ Dispersion ▼ Type

Type	▼ Frequency: 01 - Freq	Dispersion: 01_CV
the	6157.000000	0.548700
and	2465.000000	0.658276
of	2411.000000	0.757319
a	2326.000000	0.859471
is	1954.000000	0.868201
to	1680.000000	0.780321
are	1618.000000	1.086940
in	1617.000000	0.939618
that	996.000000	1.152942
they	880.000000	1.431558
it	840.000000	1.274232
water	761.000000	2.064976
animals	713.000000	1.891993
can	680.000000	1.553853
have	628.000000	1.692628
you	567.000000	1.940654
from	544.000000	1.650789
plants	536.000000	2.124250
or	529.000000	1.704293
what	516.000000	1.547638
on	474.000000	1.580419
for	471.000000	1.686723
this	422.000000	1.740056
food	402.000000	2.850878
an	393.000000	2.020543
as	391.000000	1.954979
when	390.000000	2.057010
their	389.000000	2.096211
called	386.000000	1.806971
some	370.000000	2.081900
with	341.000000	2.203422
all	323.000000	2.160366
plant	316.000000	2.689868
energy	301.000000	4.686044
1	297.000000	1.427943
be	297.000000	2.074458
2	293.000000	1.465036
at	290.000000	1.902082
earth	287.000000	2.688635
by	277.000000	2.160326
like	276.000000	2.527721
live	274.000000	2.514258
many	273.000000	1.963246
which	273.000000	2.113433

▼ Corpus COREL_SCI ▼ Frequency ▼ Dispersion ▼ Type

Type	▼ Frequency: 01 - Freq	Dispersion: 01_CV
do	268.000000	2.058796
there	263.000000	2.235166
into	259.000000	2.085096
other	255.000000	2.213006
one	251.000000	2.295064
animal	249.000000	2.505932
different	241.000000	2.272562
sun	238.000000	3.579334
air	237.000000	3.399249
these	236.000000	2.130977
how	236.000000	2.283124
make	218.000000	2.385208
up	216.000000	2.372249
about	209.000000	2.285959
has	209.000000	2.470469
its	207.000000	2.396858
grow	204.000000	2.739901
also	204.000000	2.376450
things	203.000000	3.496387
not	198.000000	2.238161
more	195.000000	2.482271
we	193.000000	4.252147
part	192.000000	2.739751
where	192.000000	2.715204
each	189.000000	2.360721
eat	182.000000	3.935942
but	181.000000	2.398949
living	178.000000	3.745595
most	177.000000	2.548879
rock	175.000000	4.860308
use	174.000000	2.574042
very	172.000000	3.006173
them	171.000000	2.685514
out	170.000000	2.747273
look	168.000000	2.369380
soil	168.000000	4.254588
answer	168.000000	1.646256
help	167.000000	2.682631
directions	166.000000	1.729780
if	165.000000	2.615972
questions	164.000000	1.661177
insects	163.000000	4.733298
leaves	162.000000	3.609455
need	158.000000	3.691797
life	157.000000	4.017784
read	154.000000	1.665975
your	152.000000	4.189534
through	152.000000	3.026419
text	150.000000	1.822360
get	150.000000	2.940933
change	148.000000	3.575505
3	146.000000	1.979662

▼ Corpus COREL_SCI ▼ Frequency ▼ Dispersion ▼ Type

Type	▼ Frequency: 01 - Freq	Dispersion: 01_CV
will	146.000000	3.221541
so	144.000000	3.006773
parts	143.000000	3.830541
made	141.000000	4.079902
seeds	138.000000	4.578555
trees	137.000000	3.904951
ocean	137.000000	5.562053
light	134.000000	5.979726
than	130.000000	3.181286
move	130.000000	4.292610
because	128.000000	2.726042
people	126.000000	3.266889
moon	126.000000	5.836398
see	124.000000	3.345925
time	120.000000	3.203165
day	119.000000	3.971125
rocks	118.000000	5.977843
changes	116.000000	3.652870
ice	116.000000	5.275472
does	114.000000	2.974748
new	113.000000	3.183080
then	113.000000	3.382198
only	113.000000	3.517349
long	112.000000	3.322141
planet	112.000000	6.373034
why	111.000000	2.816613
small	111.000000	3.585505
same	110.000000	3.358347
around	109.000000	3.652137
place	109.000000	4.180701
cold	106.000000	4.003694
body	106.000000	4.944081
ground	104.000000	3.468971
wind	103.000000	6.816269
would	102.000000	3.332235
mammals	102.000000	7.002593
form	101.000000	3.878225
two	100.000000	3.433992
flowers	100.000000	5.574919
over	100.000000	4.090167
cycle	100.000000	4.613592
land	98.000000	3.661572
habitat	96.000000	4.604125
find	96.000000	3.507878
matter	95.000000	8.075558
too	95.000000	3.604307
down	94.000000	3.637637
during	94.000000	4.104876
warm	93.000000	4.023980
fish	91.000000	5.636536
types	91.000000	4.730888
winter	90.000000	5.152565

2 Most frequent 3-grams

▼ Corpus COREL_SCI ▼ Frequency ▼ Dispersion ▼ Type

Type	▼ Frequency: 01 - Freq	Dispersio
answer the questions	160.000000	1.700766
directions read the	144.000000	1.767599
read the text	144.000000	1.767443
plants and animals	92.000000	4.081426
the text and	82.000000	2.446939
part of the	80.000000	3.925408
and answer the	77.000000	2.500059
text and answer	72.000000	2.564662
from the sun	64.000000	5.880245
the text answer	56.000000	2.911929
text answer the	56.000000	2.911929
look at the	56.000000	3.447484
live in the	52.000000	5.020675
1 what is	38.000000	4.006914
a lot of	38.000000	4.594797
what is the	36.000000	4.759578
of the plant	36.000000	6.904292
made up of	35.000000	5.124235
is called a	33.000000	5.165664
of a plant	31.000000	6.491521
2 what is	31.000000	4.408960
one of the	31.000000	6.917900
during the day	30.000000	7.041411
out of the	28.000000	6.309737
what is a	28.000000	5.079178
that live in	28.000000	6.981337
in the sky	26.000000	7.521832
parts of the	25.000000	5.515294
all living things	25.000000	7.102508
life cycle of	24.000000	6.708497
this is called	24.000000	6.379449
animals that live	24.000000	8.398348
is made up	24.000000	6.529616
in the ocean	24.000000	8.918942
in the water	24.000000	6.616501
a food chain	24.000000	8.514944
their own food	24.000000	6.899489
in the air	23.000000	5.799976
some of the	23.000000	9.030602
in the box	22.000000	5.299175
part of a	22.000000	7.486438
most of the	22.000000	7.311473
animals live in	22.000000	6.733542
the amount of	22.000000	6.252648
the animals that	22.000000	6.037052
the food chain	21.000000	8.891954
plant or animal	21.000000	10.584149
the solar system	21.000000	9.152393

▼ Corpus COREL_SCI ▼ Frequency ▼ Dispersion ▼ Type

Type	▼ Frequency: 01 - Freq	Dispersio
you can see	20.000000	6.945183
the top of	20.000000	7.541885
make their own	20.000000	7.180363
are made of	20.000000	7.483363
different types of	20.000000	7.204489
in the world	20.000000	6.262951
around the sun	19.000000	8.875305
it is the	19.000000	6.983533
from the soil	19.000000	7.593536
the sun is	19.000000	6.931209
use the words	18.000000	5.629252
is called the	18.000000	7.305766
species or types	18.000000	9.694036
an animal that	18.000000	7.338388
cycle of a	17.000000	8.235382
of the world	17.000000	7.451717
take care of	17.000000	7.658904
of the moon	17.000000	9.114732
there are many	17.000000	6.809071
what kind of	17.000000	6.518146
in the winter	17.000000	8.346872
the ground and	16.000000	6.929451
of plants and	16.000000	7.560795
the water cycle	16.000000	9.681167
energy from the	16.000000	9.416204
of the ocean	16.000000	8.482855
the greenhouse effect	15.000000	11.268722
words in the	15.000000	6.361890
or types of	15.000000	9.992099
of the flower	15.000000	11.867675
the sun and	15.000000	12.204727
this means that	15.000000	6.866146
carbon dioxide and	15.000000	6.944670
different kinds of	15.000000	7.821432
is made of	15.000000	9.144064
in the solar	15.000000	9.612297
of the water	15.000000	7.795622
animals and plants	15.000000	8.651767
draw a line	15.000000	6.174233
the words in	15.000000	6.361890
an example of	14.000000	8.389070
it can be	14.000000	7.265701
animals that are	14.000000	6.815965
all of the	14.000000	7.992676
different parts of	14.000000	7.887638
are more than	14.000000	8.336118
live on land	14.000000	11.527907
what you need	14.000000	5.902229
at the picture	14.000000	6.406353
such as a	14.000000	8.325355
is an animal	14.000000	9.377624
there are about	14.000000	9.813621

▼ Corpus COREL_SCI ▼ Frequency ▼ Dispersion ▼ Type

Type	▼ Frequency: 01 - Freq	Dispersio
of the same	14.000000	7.746772
this process is	13.000000	10.252179
a question you	13.000000	7.254327
this is the	13.000000	8.741328
what to do	13.000000	6.201046
do you think	13.000000	7.154765
living things need	13.000000	10.321822
the box to	13.000000	6.828403
side of the	13.000000	12.859934
can be found	13.000000	8.259115
animals that have	13.000000	8.961791
it is a	13.000000	9.102887
have in common	13.000000	12.021586
point to the	13.000000	6.732991
are eaten by	13.000000	13.277178
there are more	13.000000	8.765149
live in a	13.000000	10.001189
in the ground	13.000000	7.844631
3 what is	13.000000	6.521829
a life cycle	13.000000	10.882535
to make food	13.000000	9.608521
all insects have	13.000000	15.232891
away from the	13.000000	10.229091
of the ground	12.000000	8.882058
process is called	12.000000	10.448726
the end of	12.000000	9.552540
is one of	12.000000	8.515574
the questions the	12.000000	6.920348
back into the	12.000000	7.745146
the form of	12.000000	9.368101
in the form	12.000000	9.368101
do not have	12.000000	9.156411
say a question	12.000000	7.951056
are animals that	12.000000	8.162095
in the spring	12.000000	12.763913
of the plants	12.000000	9.469050
animals that eat	12.000000	13.997393
animals such as	12.000000	9.067783
lives in the	12.000000	7.718895
there are three	12.000000	7.034841
top of the	12.000000	8.907168
kinds of plants	12.000000	9.211172
the plants and	12.000000	8.463024
the remains of	12.000000	9.578278
complete the sentences	12.000000	6.392196
they are called	12.000000	10.802150
there is a	12.000000	7.212756
what part of	12.000000	7.449575
to complete the	12.000000	6.908813
which is a	12.000000	8.157489
many kinds of	12.000000	10.681292
the same kind	11.000000	9.037936

3 Most frequent 4-grams

▼ Corpus COREL_SCI ▼ Frequency ▼ Dispersion ▼ Type

Type	▼ Frequency: 01 - Freq	Dispersio
we are going to	7.000000	10.220014
1 what is a	7.000000	8.258227
the particles of matter	7.000000	17.768912
and say its name	7.000000	8.983901
the flower can make	6.000000	20.880613
from place to place	6.000000	10.426854
how can you tell	6.000000	12.155756
the chart answer the	6.000000	9.020654
things all insects have	6.000000	20.880613
then i saw the	6.000000	20.880613
named after the roman	6.000000	20.880613
are some of the	6.000000	17.483775
is a type of	6.000000	9.474835
when the sun is	6.000000	11.285919
flower can make seeds	6.000000	20.880613
dead plants and animals	6.000000	9.403469
and nutrients from the	6.000000	11.540620
life cycle is the	6.000000	12.744653
did you know that	6.000000	11.187492
process is called photosynthesis	6.000000	11.520295
the sun and the	6.000000	12.380737
insects have six legs	6.000000	11.577640
the animals that live	6.000000	16.908361
a lot of rain	6.000000	9.047287
light from the sun	6.000000	11.010134
more than half of	6.000000	11.146026
kinds of plants and	6.000000	12.137886
what kind of animal	6.000000	11.600026
chart answer the questions	6.000000	9.020654
plants and animals are	6.000000	12.105053
at the picture answer	6.000000	9.123826
during the day and	6.000000	12.416217
everywhere in the world	6.000000	11.060861
this is called the	6.000000	10.893443
of the food chain	6.000000	12.383128
we're going to learn	6.000000	10.649728
are animals that have	6.000000	11.771132
water from the soil	6.000000	14.880385
plants make their own	6.000000	12.436910
end of the stems	6.000000	14.923644
water and nutrients from	5.000000	16.235523
the roots of a	5.000000	12.295732
test what you need	5.000000	9.788891
what might happen if	5.000000	11.476939
pacific ring of fire	5.000000	20.880613

▼ Corpus COREL_SCI ▼ Frequency ▼ Dispersion ▼ Type

Type	▼ Frequency: 01 - Freq	Dispersio
a food chain shows	5.000000	12.957618
types of plants and	5.000000	13.514670
after the roman god	5.000000	20.880613
look at the animals	5.000000	9.722064
the order in which	5.000000	12.592261
an example of a	5.000000	11.281815
in the food chain	5.000000	10.515361
parts of a plant	5.000000	11.322978
a wide variety of	5.000000	14.809277
the chart then answer	5.000000	9.331109
need water to grow	5.000000	9.983421
carbon dioxide and water	5.000000	12.481646
the female part of	5.000000	16.538882
rises in the east	5.000000	12.827323
many kinds of plants	5.000000	15.546666
top of a mountain	5.000000	13.785208
plant or animal lives	5.000000	12.984152
the text and study	5.000000	10.035884
have adapted to the	5.000000	20.880613
2 what kind of	5.000000	10.908206
text and study the	5.000000	10.035884
why or why not	5.000000	9.586948
different parts of the	5.000000	9.915123
from the sun to	5.000000	10.671608
find out more about	5.000000	10.157490
they are able to	5.000000	10.196729
from flower to flower	5.000000	10.334970
of water on earth	5.000000	10.809683
the pacific ring of	5.000000	20.880613
is the name of	5.000000	11.934979
phases of the moon	5.000000	11.920529
what happens to the	5.000000	11.347394
come out of the	5.000000	11.340758
the volume of a	5.000000	20.880613
what part of the	5.000000	11.540338
the ones that interest	5.000000	12.300825
the balloon with the	5.000000	20.880613
largest planet in the	5.000000	15.463933
at the same time	5.000000	13.161981
a life cycle is	5.000000	11.153151
the temperature of the	5.000000	11.062556
take care of the	5.000000	11.520661
of plant or animal	5.000000	19.297195
is a lot of	5.000000	9.880609
oil and natural gas	5.000000	16.033140
of the same kind	5.000000	11.166477
take in carbon dioxide	5.000000	14.405454
plants and animals need	5.000000	12.154218
some of the things	5.000000	17.955685
here's how you can	5.000000	20.880613
is a force that	5.000000	16.320744
have a hard time	5.000000	14.111685

Source: #LancsBox 6.0.

▼ Corpus COREL_SCI ▼ Frequency ▼ Dispersion ▼ Type

Type	▼ Frequency: 01 - Freq	Dispersio
what happens after that	5.000000	20.880613
the north and south	5.000000	13.287230
north and south poles	5.000000	12.210072
chart then answer the	5.000000	9.331109
revolve around the sun	5.000000	15.644130
sun rises and sets	5.000000	13.579172
there is a lot	5.000000	9.880609
the parts of the	5.000000	10.212534
brightest object in the	5.000000	18.949218
ones that interest a	5.000000	12.300825
to the ground and	5.000000	11.256013
have you ever seen	5.000000	11.016563
food chain a food	5.000000	12.682745
plants that live in	5.000000	14.548175
the male part of	5.000000	16.538882
long periods of time	5.000000	13.452592
hold soil in place	5.000000	14.350426
water vapor in the	5.000000	12.328971
there are so many	5.000000	10.934200
cycle of a butterfly	5.000000	14.671126
look at the chart	5.000000	12.302016
grow into new plants	5.000000	12.032263
to look at the	5.000000	13.522206
the surface of the	5.000000	12.891655
made of rock and	5.000000	14.200270
roots of a plant	5.000000	12.295732
what would happen if	5.000000	12.318760
the sun rises and	5.000000	13.579172
so the flower can	5.000000	20.880613
the water cycle is	4.000000	18.231074
pollen is moved from	4.000000	16.048912
insects are the largest	4.000000	15.535460
lungs live on land	4.000000	20.880613
the text and look	4.000000	10.601382
you look at the	4.000000	11.658139
it is on average	4.000000	13.556159
help the plant get	4.000000	13.392762
are all examples of	4.000000	11.801387
from the soil and	4.000000	12.110316
carbon dioxide in the	4.000000	14.552938
and look at the	4.000000	10.601382
in a temperate forest	4.000000	16.240265
animals and plants that	4.000000	13.821545
stars are very far	4.000000	16.485578
is released into the	4.000000	12.115205
also known as the	4.000000	11.533887
the picture of the	4.000000	11.725986
draw a line from	4.000000	12.162534
the water in the	4.000000	11.766153
how long does it	4.000000	13.729319
the water cycle the	4.000000	12.029889
the flow of water	4.000000	14.408347

Source: #LancsBox 6.0.

Appendix N Most frequent word classes and n-grams: *Neighborhood* (T1)

1 Nouns

#LancsBox 6.0

KWIC | GraphColl | Whelk | Words | N

Corpora | Words: Corpus 4 - Animals | Ngrams: Corpus 4 - Animals | Words: Corpus 5 - Neighborhood

Search

▼ Corpus Corpus 5 - Neighborhood ▼ Frequency ▼ Dispersion ▼ Lem

Lemma	Frequency: 01 - Freq	Dispersion: 01_CV
people_n	96.000000	1.155172
city_n	61.000000	1.674136
place_n	54.000000	1.458141
map_n	47.000000	1.665468
community_n	47.000000	1.978356
school_n	35.000000	1.736022
home_n	26.000000	1.950122
world_n	21.000000	4.300135
building_n	21.000000	2.209308
area_n	20.000000	2.627697
library_n	19.000000	2.029869
question_n	18.000000	1.481799
street_n	17.000000	2.754022
neighborhood_n	16.000000	3.437658
photo_n	16.000000	1.639778
park_n	16.000000	2.198210
land_n	16.000000	3.516176
market_n	16.000000	3.553789
direction_n	15.000000	3.749130
text_n	15.000000	1.697263
house_n	15.000000	1.851981
population_n	14.000000	4.103200
town_n	14.000000	3.358024
police_n	13.000000	2.833325
country_n	13.000000	3.556306
farmer_n	12.000000	2.931712
thing_n	12.000000	2.302623
station_n	12.000000	3.967828
food_n	12.000000	4.095858
theater_n	12.000000	2.851192
language_n	11.000000	6.480741
service_n	11.000000	4.327246
farm_n	11.000000	2.611510
store_n	11.000000	3.042920
fire_n	10.000000	3.301131
office_n	10.000000	2.727099
grid_n	10.000000	4.075438
suburb_n	9.000000	4.255749

beach_n	9.000000	4.841045
part_n	9.000000	4.191488
study_n	9.000000	2.371821
west_n	9.000000	2.195480
kind_n	9.000000	3.161309
type_n	8.000000	2.984901
family_n	8.000000	4.313630
letter_n	8.000000	3.273161
draw_n	8.000000	2.523050
hospital_n	8.000000	2.996759
gas_n	8.000000	4.308893
subway_n	8.000000	6.291199
wheat_n	8.000000	6.480741
reason_n	7.000000	3.797126
south_n	7.000000	2.390209
lot_n	7.000000	3.228685
north_n	7.000000	2.348606
job_n	7.000000	3.307458
united_n	7.000000	4.856558
room_n	7.000000	4.450528
book_n	7.000000	3.700441
restroom_n	7.000000	3.668712
officer_n	7.000000	3.673915
business_n	7.000000	2.530599
dog_n	7.000000	6.205536
states_n	7.000000	4.856558
bus_n	7.000000	4.020677
key_n	6.000000	5.283545
municipality_n	6.000000	6.480741
line_n	6.000000	3.562835
transportation_n	6.000000	3.407807
center_n	6.000000	3.447100
shop_n	6.000000	3.199069
car_n	6.000000	3.053806
space_n	6.000000	3.629254
picture_n	6.000000	3.120709
road_n	6.000000	3.098090
st._n	6.000000	6.480741
water_n	6.000000	3.281608
time_n	6.000000	2.975803
movie_n	6.000000	4.007088
east_n	6.000000	2.534741
other_n	6.000000	3.941405
route_n	6.000000	5.218327
name_n	6.000000	3.002869
supermarket_n	6.000000	3.135554
example_n	6.000000	2.744286
good_n	6.000000	2.848694
information_n	6.000000	3.263184
word_n	5.000000	3.248924
citizen_n	5.000000	6.241838
post_n	5.000000	3.524504

2 Verbs

Lemma	Frequency: 01 - Freq	Dispersion: 01_CV
eat_v	6.000000	3.404784
draw_v	6.000000	2.853700
build_v	6.000000	2.844863
play_v	6.000000	3.706170
fill_v	5.000000	3.334110
put_v	5.000000	3.087660
bring_v	5.000000	2.973226
spread_v	5.000000	4.230458
close_v	4.000000	4.753017
call_v	4.000000	3.968625
happen_v	4.000000	3.873753
complete_v	4.000000	3.873459
follow_v	4.000000	4.206770
walk_v	4.000000	3.342583
store_v	4.000000	3.852456
locate_v	4.000000	4.796545
set_v	3.000000	3.750017
label_v	3.000000	4.015561
learn_v	3.000000	5.584701
relax_v	3.000000	4.691716
raise_v	3.000000	4.825343
think_v	3.000000	3.804795
map_v	3.000000	4.326258
place_v	3.000000	4.811960
mean_v	3.000000	4.196332
add_v	3.000000	4.584275
deliver_v	3.000000	3.927540
ride_v	3.000000	3.719685
clean_v	3.000000	4.235320
understand_v	3.000000	3.698807
describe_v	3.000000	3.765813
say_v	3.000000	3.712091
rise_v	2.000000	6.480741
hike_v	2.000000	4.549999
include_v	2.000000	4.784233
love_v	2.000000	4.967765
crowd_v	2.000000	4.694779
harm_v	2.000000	6.480741
provide_v	2.000000	5.788452
teach_v	2.000000	4.910974
stop_v	2.000000	4.888998
ask_v	2.000000	4.613030
roll_v	2.000000	4.623082
beach_v	2.000000	6.480741
check_v	2.000000	4.611891
pick_v	2.000000	6.480741
fall_v	2.000000	4.655598
shop_v	2.000000	4.770590
begin_v	2.000000	4.632858
continue_v	2.000000	6.480741
immigrate_v	2.000000	6.480741
reach_v	2.000000	4.585664

Lemma	Frequency: 01 - Freq	Dispersion: 01_CV
decide_v	2.000000	6.480741
choose_v	2.000000	4.854421
honey_v	2.000000	4.812459
buy_v	2.000000	4.529301
pump_v	2.000000	6.480741
plan_v	2.000000	4.907404
catch_v	2.000000	4.928098
connect_v	2.000000	4.555975
enjoy_v	2.000000	4.533300
write_v	2.000000	6.480741
shop-rain_v	1.000000	6.480741
cut_v	1.000000	6.480741
earn_v	1.000000	6.480741
fish_v	1.000000	6.480741
hurt_v	1.000000	6.480741
sleep_v	1.000000	6.480741
expand_v	1.000000	6.480741
shine_v	1.000000	6.480741
type_v	1.000000	6.480741
protect_v	1.000000	6.480741
refer_v	1.000000	6.480741
start_v	1.000000	6.480741
border_v	1.000000	6.480741
practice_v	1.000000	6.480741
park_v	1.000000	6.480741
cross_v	1.000000	6.480741
affect_v	1.000000	6.480741
imagine_v	1.000000	6.480741
flood_v	1.000000	6.480741
cool_v	1.000000	6.480741
wash_v	1.000000	6.480741
arrive_v	1.000000	6.480741
lead_v	1.000000	6.480741
lie_v	1.000000	6.480741
pursue_v	1.000000	6.480741
withdraw_v	1.000000	6.480741
share_v	1.000000	6.480741
house_v	1.000000	6.480741
steal_v	1.000000	6.480741
represent_v	1.000000	6.480741
agree_v	1.000000	6.480741
stay_v	1.000000	6.480741
collect_v	1.000000	6.480741
bake_v	1.000000	6.480741
list_v	1.000000	6.480741
send_v	1.000000	6.480741
divide_v	1.000000	6.480741
cost_v	1.000000	6.480741
vote_v	1.000000	6.480741
people_v	1.000000	6.480741
slow_v	1.000000	6.480741
leave_v	1.000000	6.480741

3 Adjectives

Lemma	Frequency: 01 - Freq	Dispersion: 01_CV
many_adj	35.000000	1.320839
rural_adj	17.000000	2.682700
different_adj	16.000000	1.853968
other_adj	15.000000	2.514234
more_adj	12.000000	3.357533
few_adj	11.000000	2.267644
large_adj	10.000000	4.034386
such_adj	9.000000	2.581147
urban_adj	9.000000	3.669938
good_adj	9.000000	3.879512
small_adj	9.000000	2.988021
big_adj	8.000000	2.556353
suburban_adj	6.000000	3.960460
public_adj	6.000000	3.034805
most_adj	6.000000	3.023783
natural_adj	6.000000	4.540614
new_adj	6.000000	2.829009
open_adj	5.000000	3.880316
busy_adj	5.000000	2.906156
safe_adj	5.000000	4.645469
local_adj	5.000000	3.772602
close_adj	5.000000	3.795491
official_adj	5.000000	6.480741
special_adj	4.000000	3.993565
near_adj	4.000000	3.850986
pet_adj	4.000000	6.480741
high_adj	4.000000	4.200514
easy_adj	3.000000	3.686901
short_adj	3.000000	4.697126
2nd_adj	3.000000	4.116360
great_adj	3.000000	4.849667
fast_adj	3.000000	4.784233
long_adj	3.000000	3.742308
healthy_adj	3.000000	4.528564
particular_adj	3.000000	3.690487
own_adj	3.000000	5.215834
next_adj	3.000000	3.815110
hard_adj	3.000000	4.902408
human_adj	3.000000	6.480741
important_adj	3.000000	4.159700
much_adj	3.000000	4.641856
same_adj	2.000000	6.480741
diverse_adj	2.000000	6.480741
grassy_adj	2.000000	4.694779
less_adj	2.000000	5.099005
available_adj	2.000000	6.480741
flat_adj	2.000000	5.349533
top_adj	2.000000	6.480741
common_adj	2.000000	4.589926
correct_adj	2.000000	4.554137
steep_adj	2.000000	6.480741
tasty_adj	2.000000	4.529301

Lemma	Frequency: 01 - Freq	Dispersion: 01_CV
regular_adj	2.000000	5.450119
first_adj	2.000000	4.645644
nearby_adj	2.000000	4.583544
empty_adj	2.000000	4.643188
similar_adj	2.000000	4.536706
real_adj	2.000000	6.480741
northeast_adj	2.000000	6.480741
english_adj	2.000000	6.480741
old_adj	2.000000	4.967765
elderly_adj	1.000000	6.480741
dynamic_adj	1.000000	6.480741
wonderful_adj	1.000000	6.480741
photo_adj	1.000000	6.480741
6th_adj	1.000000	6.480741
lovely_adj	1.000000	6.480741
federal_adj	1.000000	6.480741
include_adj	1.000000	6.480741
daily_adj	1.000000	6.480741
hilly_adj	1.000000	6.480741
name_adj	1.000000	6.480741
middle_adj	1.000000	6.480741
left_adj	1.000000	6.480741
helpful_adj	1.000000	6.480741
three-day_adj	1.000000	6.480741
warm_adj	1.000000	6.480741
polite_adj	1.000000	6.480741
popular_adj	1.000000	6.480741
enough_adj	1.000000	6.480741
a4_adj	1.000000	6.480741
valuable_adj	1.000000	6.480741
complete_adj	1.000000	6.480741
create_adj	1.000000	6.480741
audio_adj	1.000000	6.480741
librarian_adj	1.000000	6.480741
korean_adj	1.000000	6.480741
clear_adj	1.000000	6.480741
open-air_adj	1.000000	6.480741
low_adj	1.000000	6.480741
11th_adj	1.000000	6.480741
fresh_adj	1.000000	6.480741
several_adj	1.000000	6.480741
east_adj	1.000000	6.480741
afraid_adj	1.000000	6.480741
harmful_adj	1.000000	6.480741
1800s_adj	1.000000	6.480741
municipal_adj	1.000000	6.480741
fair_adj	1.000000	6.480741
dangerous_adj	1.000000	6.480741
animal_adj	1.000000	6.480741
sick_adj	1.000000	6.480741
midwest_adj	1.000000	6.480741
fortunate_adj	1.000000	6.480741

4 Adverbs

Lemma	Frequency: 01 - Freq	Dispersion: 01_CV
more_adv	6.000000	3.300219
so_adv	6.000000	3.161966
together_adv	6.000000	3.005613
below_adv	5.000000	2.856532
too_adv	5.000000	4.216524
sometimes_adv	5.000000	2.936133
still_adv	4.000000	3.199796
close_adv	4.000000	3.767509
usually_adv	4.000000	3.589071
south_adv	4.000000	3.687625
just_adv	4.000000	3.169457
far_adv	3.000000	3.657377
all_adv	3.000000	3.833314
often_adv	3.000000	4.190150
north_adv	3.000000	3.725278
most_adv	3.000000	4.309134
apart_adv	3.000000	3.708253
home_adv	3.000000	4.072224
else_adv	3.000000	3.742013
down_adv	3.000000	3.741092
as_adv	3.000000	4.256767
around_adv	2.000000	4.651826
now_adv	2.000000	5.117278
instead_adv	2.000000	4.585664
ever_adv	2.000000	4.562145
long_adv	2.000000	4.651826
even_adv	2.000000	4.551062
only_adv	2.000000	4.529191
less_adv	1.000000	6.480741
once_adv	1.000000	6.480741
closer_adv	1.000000	6.480741
farther_adv	1.000000	6.480741
straight_adv	1.000000	6.480741
underground_adv	1.000000	6.480741
however_adv	1.000000	6.480741
aloud_adv	1.000000	6.480741
ill_adv	1.000000	6.480741
always_adv	1.000000	6.480741
above_adv	1.000000	6.480741
away_adv	1.000000	6.480741
east_adv	1.000000	6.480741
early_adv	1.000000	6.480741
ago_adv	1.000000	6.480741
hard_adv	1.000000	6.480741
equally_adv	1.000000	6.480741
gently_adv	1.000000	6.480741
sure_adv	1.000000	6.480741
differently_adv	1.000000	6.480741
back_adv	1.000000	6.480741
item_adv	1.000000	6.480741
closely_adv	1.000000	6.480741
lot_adv	1.000000	6.480741

5 3-grams

▼ Corpus Corpus 5 - Neighborhood ▼ Frequency ▼ Dispersion ▼ Type

Type	▼ Frequency: 01 - Freq	Dispersion: (
answer the questions	18.000000	1.496661
then answer the	15.000000	1.710069
study the photo	13.000000	1.753243
this is a	13.000000	6.480741
read the text	13.000000	1.676720
the photo then	11.000000	1.874867
photo then answer	10.000000	1.943026
and study the	8.000000	2.265642
text and study	8.000000	2.265642
the text and	8.000000	2.265642
on the map	7.000000	2.825933
map of a	7.000000	2.334853
there is a	6.000000	2.841878
a lot of	6.000000	3.687286
the natural world	6.000000	4.541336
the united states	6.000000	6.004359
live in a	6.000000	3.020852
a map of	5.000000	3.180265
a farmers market	5.000000	3.993006
look at this	5.000000	2.843496
you want to	5.000000	3.340437
in the city	5.000000	4.372773
at a farmers	4.000000	4.169114
their home countries	4.000000	5.459973
of the natural	4.000000	4.541336
is a lot	4.000000	3.756763
in a community	4.000000	4.528115
you live in	4.000000	3.754058
a good citizen	4.000000	6.480741
the land in	4.000000	5.640306
lot of open	4.000000	4.129518
to get to	4.000000	3.687060
at this map	4.000000	3.171967
to the united	4.000000	6.480741
do you know	4.000000	3.442547
fruits and vegetables	4.000000	4.532792
there are many	4.000000	3.882068
a rural community	4.000000	4.417481
west east south	4.000000	3.177665
part of the	4.000000	6.480741
you have a	3.000000	4.239334
in your community	3.000000	4.465660
in a city	3.000000	3.743347
the human world	3.000000	6.480741
north west east	3.000000	3.732227
transportation to get	3.000000	3.717807
plants and animals	3.000000	4.850532
a big city	3.000000	4.122050
people move for	3.000000	6.480741
places in the	3.000000	4.536015
of open space	3.000000	5.302537
be used to	3.000000	6.480741

▼ Corpus Corpus 5 - Neighborhood ▼ Frequency ▼ Dispersion ▼ Type

Type	▼ Frequency: 01 - Freq	Dispersion: (
the map of	3.000000	3.804387
want to be	3.000000	6.480741
at the park	3.000000	3.990864
go to the	3.000000	3.850195
of the city	3.000000	6.480741
public transportation to	3.000000	3.717807
used to make	3.000000	6.480741
rural community has	3.000000	4.836658
the text study	3.000000	3.658687
this map of	3.000000	3.681072
work in the	3.000000	6.480741
you can help	3.000000	6.480741
the questions urban	3.000000	4.033184
a map can	3.000000	3.728048
what kind of	3.000000	4.598685
something that is	3.000000	6.084891
how do you	3.000000	4.161866
community is a	3.000000	4.648164
you need to	3.000000	4.375590
draw a line	3.000000	4.111065
a community is	3.000000	4.610116
in their home	3.000000	6.480741
come to the	3.000000	4.530629
of a school	3.000000	4.066248
is in it	3.000000	6.480741
text study the	3.000000	3.658687
different types of	3.000000	4.878912
community service is	3.000000	6.480741
people come to	3.000000	4.530629
in a place	3.000000	4.548888
have to drive	3.000000	4.721071
different from a	2.000000	4.539790
being a good	2.000000	6.480741
in the order	2.000000	6.480741
of a place	2.000000	5.295137
gym restroom classroom	2.000000	4.556910
over the world	2.000000	4.657642
live very close	2.000000	6.480741
be near a	2.000000	6.480741
to be close	2.000000	6.480741
choose to live	2.000000	4.865156
is part of	2.000000	6.480741
where many people	2.000000	5.044036
may need more	2.000000	6.480741
places on the	2.000000	6.480741
to find your	2.000000	4.556910
the text then	2.000000	4.756600
came here to	2.000000	6.480741
supermarket movie theater	2.000000	4.529706
is much to	2.000000	6.480741
many kinds of	2.000000	6.480741
have homes and	2.000000	6.480741

6 4-grams

▼ Corpus Corpus 5 - Neighborhood ▼ Frequency ▼ Dispersion ▼ Type

Type	▼ Frequency: 01 - Freq	Dispersion: (
more fewer pet dogs	2.000000	6.480741
are part of the	2.000000	6.480741
a map of a	2.000000	4.593171
out the following information	2.000000	6.480741
path on the map	2.000000	4.713106
these places in the	2.000000	6.480741
places you would find	2.000000	6.480741
room library gym restroom	2.000000	4.557604
they are more spread	2.000000	4.617756
the number of people	2.000000	6.480741
can be used to	2.000000	6.480741
map of a school	2.000000	4.557604
work in the city	2.000000	6.480741
kinds of food in	2.000000	6.480741
might want to be	2.000000	6.480741
office science room library	2.000000	4.557604
beaches are great places	2.000000	6.480741
from place to place	2.000000	4.825598
they might want to	2.000000	6.480741
what direction will you	2.000000	6.480741
pictures of places you	2.000000	6.480741
the text then answer	2.000000	4.758423
do you live in	2.000000	6.480741
supermarket movie theater drugstore	2.000000	4.529746
library gym restroom classroom	2.000000	4.557604
to be close to	2.000000	6.480741
a map of your	2.000000	4.553924
of places you would	2.000000	6.480741
is a group of	2.000000	5.405054
that is part of	2.000000	6.480741
is the number of	2.000000	6.480741
a community is a	2.000000	5.070230
many kinds of food	2.000000	6.480741
kind of community do	2.000000	6.480741
a good place to	2.000000	6.480741
all over the world	2.000000	4.658598
movie theater drugstore hospital	2.000000	4.529746
to find your way	2.000000	4.557604
give information about the	2.000000	4.533624
population is the number	2.000000	6.480741
direction will you go	2.000000	6.480741
you would find in	2.000000	6.480741
of the human world	2.000000	6.480741
population population is the	2.000000	6.480741
being a good citizen	2.000000	6.480741
science room library gym	2.000000	4.557604
many people live close	2.000000	5.048539
for many reasons they	2.000000	6.480741
there is much to	2.000000	6.480741
is part of the	2.000000	6.480741
but they are more	2.000000	4.617756
room office science room	2.000000	4.557604

▼ Corpus Corpus 5 - Neighborhood ▼ Frequency ▼ Dispersion ▼ Type

Type	▼ Frequency: 01 - Freq	Dispersion: (
information about your nearest	2.000000	6.480741
are great places to	2.000000	6.480741
the following information about	2.000000	6.480741
want to be close	2.000000	6.480741
food in the city	2.000000	6.480741
the word box below	2.000000	4.557396
they need to go	2.000000	4.556910
looking at a map	2.000000	6.480741
the united states from	2.000000	6.480741
at the map of	2.000000	4.600775
part of the natural	2.000000	6.480741
what type of community	2.000000	6.480741
to go from the	2.000000	6.480741
fewer pet dogs than	2.000000	6.480741
would find in a	2.000000	6.480741
places in the order	2.000000	6.480741
the same color as	2.000000	6.480741
the pictures of places	2.000000	6.480741
community do you live	2.000000	6.480741
directions read the text	2.000000	4.566136
the names of the	2.000000	6.480741
find out the following	2.000000	6.480741
from all over the	2.000000	4.658598
which item sold the	2.000000	6.480741
text then answer the	2.000000	4.758423
reading terminal market is	2.000000	6.480741
live in a place	2.000000	4.922666
and you want to	2.000000	5.459686
to put out fires	2.000000	5.071809
look at the map	2.000000	4.600775
people choose to live	2.000000	4.870722
lunch room office science	2.000000	4.557604
are more spread out	2.000000	4.617756
of community do you	2.000000	6.480741
the reading terminal market	2.000000	6.480741
something that is part	2.000000	6.480741
to walk and take	1.000000	6.480741
located korean war memorial	1.000000	6.480741
the natural world the	1.000000	6.480741
animals at the beaches	1.000000	6.480741
our house is called	1.000000	6.480741
together and has something	1.000000	6.480741
pool park mall gas	1.000000	6.480741
there are many teachers	1.000000	6.480741
and it is a	1.000000	6.480741
the middle square the	1.000000	6.480741
minutes but on the	1.000000	6.480741
right next to in	1.000000	6.480741
the questions write about	1.000000	6.480741
are crowded with people	1.000000	6.480741
directions use the word	1.000000	6.480741
like a train it	1.000000	6.480741

Source: #LancsBox 6.0.

Appendix O Most frequent word classes and n-grams: *Animals* (T2)

1 Nouns

▼ Corpus Corpus 4 - Animals ▼ Frequency ▼ Dispersion ▼ Lemma

Lemma	Frequency: 01 - Freq	Dispersion: 01_CV
animal_n	165.000000	1.131762
insect_n	61.000000	2.611211
mammal_n	48.000000	3.500620
bird_n	44.000000	3.173543
fish_n	43.000000	3.338003
water_n	41.000000	2.527510
type_n	39.000000	2.900862
body_n	38.000000	2.121597
group_n	38.000000	2.817308
snake_n	34.000000	4.529454
wing_n	34.000000	2.494012
leg_n	33.000000	2.589348
backbone_n	33.000000	2.710849
specie_n	32.000000	3.055625
part_n	30.000000	2.181851
cat_n	30.000000	2.716561
vertebrate_n	28.000000	4.297410
mollusk_n	28.000000	5.753066
egg_n	26.000000	2.808295
spider_n	25.000000	3.338940
baby_n	25.000000	3.557181
reptile_n	24.000000	4.608601
kind_n	23.000000	2.996490
lion_n	21.000000	3.673580
people_n	21.000000	3.341577
turtle_n	21.000000	3.544087
question_n	19.000000	2.456059
frog_n	19.000000	5.038150
night_n	18.000000	6.619612
gill_n	17.000000	6.001531
world_n	17.000000	2.740542
day_n	17.000000	4.531817
name_n	17.000000	3.007919
direction_n	16.000000	2.425315
feather_n	16.000000	3.565113
land_n	16.000000	3.558137
bear_n	16.000000	3.334622
dog_n	16.000000	3.038865
word_n	15.000000	2.552668
circle_n	15.000000	2.316115
adult_n	15.000000	2.836525
human_n	15.000000	2.971262
alligator_n	14.000000	5.007004
bee_n	14.000000	4.487166

▼ Corpus Corpus 4 - Animals ▼ Frequency ▼ Dispersion ▼ Lemma

Lemma	Frequency: 01 - Freq	Dispersion: 01_CV
invertebrate_n	14.000000	6.206184
lizard_n	14.000000	4.276279
rodent_n	13.000000	6.768086
bat_n	13.000000	4.161800
skin_n	13.000000	3.517140
shell_n	13.000000	5.342769
year_n	13.000000	4.559814
pet_n	13.000000	5.491196
bone_n	12.000000	4.064002
tooth_n	12.000000	6.382613
food_n	12.000000	3.505076
plant_n	12.000000	3.399351
pig_n	12.000000	3.967263
grasshopper_n	11.000000	3.535372
eye_n	11.000000	4.385227
arachnid_n	11.000000	6.130124
text_n	11.000000	2.735985
male_n	11.000000	7.874008
squirrel_n	11.000000	5.526754
amphibian_n	10.000000	4.303706
hair_n	10.000000	3.925714
species_n	10.000000	4.563279
shark_n	10.000000	4.691760
butterfly_n	10.000000	5.394300
prey_n	10.000000	5.298124
predator_n	10.000000	6.368418
ant_n	10.000000	4.067076
air_n	10.000000	4.629732
tail_n	10.000000	4.394651
toad_n	9.000000	5.574608
family_n	9.000000	4.281762
mouth_n	9.000000	5.676406
way_n	9.000000	3.243398
squid_n	9.000000	4.717148
job_n	9.000000	6.007568
scientist_n	9.000000	4.151377
scale_n	9.000000	4.249740
earth_n	9.000000	3.673948
zoo_n	9.000000	3.750756
elephant_n	9.000000	5.341375
parent_n	9.000000	4.265205
color_n	9.000000	4.368574
answer_n	9.000000	2.748506
female_n	9.000000	5.264348
mite_n	9.000000	4.477229
sea_n	9.000000	6.881587
oxygen_n	9.000000	7.874008
wasp_n	8.000000	5.731936
draw_n	8.000000	3.298001
birds_n	8.000000	4.243344
tiger_n	8.000000	4.490316

2 Verbs

Lemma	Frequency: 01 - Freq	Dispersion: 01_CV
be_v	494.000000	0.619691
have_v	172.000000	0.927506
live_v	77.000000	1.664547
can_v	61.000000	1.578228
do_v	55.000000	1.509275
call_v	48.000000	2.046402
use_v	35.000000	2.171928
fly_v	32.000000	3.024409
know_v	27.000000	1.983511
look_v	25.000000	2.479805
help_v	23.000000	3.111343
make_v	22.000000	2.159102
find_v	22.000000	2.552927
eat_v	20.000000	3.167151
protect_v	15.000000	3.571161
lay_v	15.000000	3.708034
breathe_v	15.000000	5.001170
relate_v	15.000000	3.064964
keep_v	14.000000	3.012413
get_v	13.000000	3.748312
move_v	13.000000	3.602592
belong_v	12.000000	3.566554
see_v	12.000000	3.437997
read_v	11.000000	2.735985
grow_v	10.000000	3.353622
take_v	10.000000	4.399693
feed_v	9.000000	3.013807
come_v	9.000000	4.205209
include_v	8.000000	3.167440
stay_v	8.000000	4.776290
kill_v	8.000000	5.667242
cover_v	8.000000	4.648049
will_v	8.000000	4.238017
would_v	7.000000	4.743155
develop_v	7.000000	4.369112
mean_v	7.000000	3.987177
answer_v	7.000000	3.446948
swim_v	7.000000	4.295867
change_v	7.000000	5.441648
sleep_v	6.000000	7.874008
produce_v	6.000000	4.367613
may_v	6.000000	4.326315
raise_v	6.000000	4.004577
build_v	6.000000	4.346900
stand_v	6.000000	5.730655
spend_v	6.000000	4.369345
hunt_v	6.000000	4.739057
like_v	6.000000	3.570919
run_v	5.000000	4.022780
feather_v	5.000000	6.134270
want_v	5.000000	4.144130

think_v	5.000000	4.490918
become_v	5.000000	4.748502
flap_v	5.000000	6.352176
need_v	5.000000	3.868921
name_v	5.000000	4.429884
walk_v	5.000000	5.085278
depend_v	4.000000	6.773954
lie_v	4.000000	4.719149
share_v	4.000000	4.302980
hold_v	4.000000	5.181166
attach_v	4.000000	6.159911
discover_v	4.000000	4.787591
play_v	4.000000	6.314132
say_v	4.000000	4.195455
teach_v	3.000000	7.166393
prefer_v	3.000000	5.556596
pull_v	3.000000	6.313527
might_v	3.000000	5.736927
gnaw_v	3.000000	5.541855
blend_v	3.000000	5.157776
happen_v	3.000000	6.385428
absorb_v	3.000000	7.874008
spread_v	3.000000	5.367048
leave_v	3.000000	5.792547
wing_v	3.000000	5.269864
tell_v	3.000000	5.186752
die_v	3.000000	5.427376
curve_v	3.000000	5.183846
catch_v	3.000000	4.561272
go_v	3.000000	5.589054
jump_v	3.000000	5.139789
lose_v	3.000000	5.728974
enjoy_v	3.000000	5.671599
cub_v	3.000000	5.497588
hide_v	3.000000	6.949908
tame_v	2.000000	5.671721
cut_v	2.000000	5.696681
attack_v	2.000000	5.734237
groom_v	2.000000	7.874008
allow_v	2.000000	5.694980
put_v	2.000000	5.580579
threaten_v	2.000000	5.750408
visit_v	2.000000	6.154277
could_v	2.000000	6.065360
create_v	2.000000	7.874008
show_v	2.000000	5.945434
warm_v	2.000000	5.534137
light_v	2.000000	5.526612
hatch_v	2.000000	7.874008
ask_v	2.000000	6.065360
adapt_v	2.000000	5.739112

3 Adjectives

extinct_adj	4.000000	5.310473
old_adj	4.000000	4.730935
yellow_adj	4.000000	4.810459
social_adj	3.000000	7.194395
bony_adj	3.000000	5.687467
light_adj	3.000000	6.426048
geometric_adj	3.000000	7.874008
adult_adj	3.000000	6.167223
less_adj	3.000000	5.306208
warm_adj	3.000000	5.522691
moist_adj	3.000000	5.334963
lay_adj	3.000000	7.874008
white_adj	3.000000	6.368446
great_adj	3.000000	5.834744
male_adj	3.000000	6.889421
scaly_adj	3.000000	6.765767
north_adj	3.000000	5.249496
scientific_adj	3.000000	5.575971
tiny_adj	3.000000	5.113099
front_adj	3.000000	5.536374
live_adj	3.000000	5.527186
unique_adj	3.000000	4.632140
black_adj	3.000000	5.641858
tame_adj	3.000000	7.028691
important_adj	3.000000	5.537600
least_adj	3.000000	7.051920
rough_adj	3.000000	6.880588
invertebrate_adj	2.000000	7.874008
easy_adj	2.000000	5.740785
coral_adj	2.000000	7.874008
shy_adj	2.000000	6.302135
shallow_adj	2.000000	6.349721
short_adj	2.000000	7.874008
eight-legged_adj	2.000000	5.665217
safe_adj	2.000000	6.019645
latin_adj	2.000000	5.529784
https://kids.britannica.com/kids/br...	2.000000	5.945368
intelligent_adj	2.000000	7.874008
simple_adj	2.000000	6.101978
half_adj	2.000000	7.874008
hollow_adj	2.000000	7.874008
tail_adj	2.000000	7.874008
enough_adj	2.000000	6.369364
thin_adj	2.000000	5.598010
second_adj	2.000000	5.965822
huge_adj	2.000000	7.874008
lizardlike_adj	2.000000	7.874008
cartilaginous_adj	2.000000	7.874008
slender_adj	2.000000	6.467185
injured_adj	2.000000	7.874008
slow-moving_adj	2.000000	5.544691
female_adj	2.000000	7.874008

Lemma	Frequency: BY Freq	Dispersion: BY_DP
first_adj	2.000000	5.544581
shell_adj	2.000000	7.874008
dangerous_adj	2.000000	5.528601
fierce_adj	2.000000	6.071160
full_adj	2.000000	5.522681
tall_adj	2.000000	5.534825
immature_adj	2.000000	7.874008
flexible_adj	2.000000	7.874008
unusual_adj	2.000000	5.600169
heavy_adj	2.000000	5.945368
single_adj	2.000000	5.769604
solid_adj	2.000000	5.769604
main_adj	2.000000	6.844472
thick_adj	2.000000	7.874008
possible_adj	2.000000	6.084300
colorful_adj	2.000000	6.262540
red_adj	2.000000	5.750408
weak_adj	2.000000	7.874008
bright_adj	2.000000	5.532169
flightless_adj	2.000000	7.874008
fawn_adj	2.000000	6.067242
much_adj	2.000000	7.874008
spotted_adj	2.000000	7.874008
stiff_adj	2.000000	5.532663
high-speed_adj	1.000000	7.874008
orange_adj	1.000000	7.874008
reptile_adj	1.000000	7.874008
inky_adj	1.000000	7.874008
impossible_adj	1.000000	7.874008
digestive_adj	1.000000	7.874008
walkingstick_adj	1.000000	7.874008
stout_adj	1.000000	7.874008
america—the_adj	1.000000	7.874008
northern_adj	1.000000	7.874008
terrifying_adj	1.000000	7.874008
feathery_adj	1.000000	7.874008
alarming_adj	1.000000	7.874008
general_adj	1.000000	7.874008
open_adj	1.000000	7.874008
harmless_adj	1.000000	7.874008
invisible_adj	1.000000	7.874008
mollusk's_adj	1.000000	7.874008
pink_adj	1.000000	7.874008
eel_adj	1.000000	7.874008
sweet_adj	1.000000	7.874008
2ndg_adj	1.000000	7.874008
fuzzy_adj	1.000000	7.874008
certain_adj	1.000000	7.874008
busy_adj	1.000000	7.874008
wide_adj	1.000000	7.874008
blooded_adj	1.000000	7.874008
famous_adj	1.000000	7.874008

4 Adverbs

Lemma	Frequency: 01 - Freq	Dispersion: 01_CV
always_adv	2.000000	5.533568
rather_adv	2.000000	6.467185
yes_adv	2.000000	5.533423
ago_adv	2.000000	6.861023
either_adv	2.000000	5.750354
on_adv	2.000000	6.475655
really_adv	2.000000	5.648537
right_adv	2.000000	5.580579
quite_adv	2.000000	5.555416
fully_adv	2.000000	6.867398
easily_adv	2.000000	5.739112
still_adv	1.000000	7.874008
far_adv	1.000000	7.874008
below_adv	1.000000	7.874008
faster_adv	1.000000	7.874008
anywhere_adv	1.000000	7.874008
normally_adv	1.000000	7.874008
longer_adv	1.000000	7.874008
short_adv	1.000000	7.874008
probably_adv	1.000000	7.874008
nearly_adv	1.000000	7.874008
awake_adv	1.000000	7.874008
freely_adv	1.000000	7.874008
closer_adv	1.000000	7.874008
enough_adv	1.000000	7.874008
some_adv	1.000000	7.874008
underwater_adv	1.000000	7.874008
deep_adv	1.000000	7.874008
in_adv	1.000000	7.874008
alike_adv	1.000000	7.874008
no_adv	1.000000	7.874008
exactly_adv	1.000000	7.874008
highly_adv	1.000000	7.874008
upside_adv	1.000000	7.874008
specially_adv	1.000000	7.874008
completely_adv	1.000000	7.874008
typically_adv	1.000000	7.874008
again_adv	1.000000	7.874008
tightly_adv	1.000000	7.874008
brightly_adv	1.000000	7.874008
worldwide_adv	1.000000	7.874008
sure_adv	1.000000	7.874008
nevertheless_adv	1.000000	7.874008
off_adv	1.000000	7.874008
home_adv	1.000000	7.874008
else_adv	1.000000	7.874008
truly_adv	1.000000	7.874008
slowly_adv	1.000000	7.874008
fairly_adv	1.000000	7.874008
asleep_adv	1.000000	7.874008
back_adv	1.000000	7.874008
partly_adv	1.000000	7.874008

5 3-grams

Type	▼ Frequency: 01 - Freq	Dispersion: 01
species or types	18.000000	3.562577
answer the questions	16.000000	2.242937
or types of	15.000000	3.679384
there are about	14.000000	3.609469
animals that are	11.000000	2.708282
read the text	11.000000	2.752182
there are more	10.000000	3.896570
are more than	10.000000	3.896570
the animals that	10.000000	2.821383
directions read the	10.000000	2.946004
that live in	9.000000	4.347907
species or kinds	9.000000	3.456406
during the day	8.000000	6.639977
or kinds of	8.000000	3.684578
are animals that	8.000000	3.163193
live in the	8.000000	3.373748
live on land	8.000000	5.488024
belong to the	7.000000	5.105445
known for their	7.000000	3.618910
are insects that	7.000000	6.340722
animals that have	7.000000	3.514612
in the world	7.000000	3.358339
and answer the	7.000000	3.453719
means that they	6.000000	3.974373
an animal that	6.000000	3.489321
are the largest	6.000000	5.055516
have a backbone	6.000000	4.159412
the text and	6.000000	3.574719
text and answer	6.000000	3.574719
group of animals	6.000000	5.055647
what is a	6.000000	5.243389
part of the	6.000000	4.689009
you have a	5.000000	6.148371
in the wild	5.000000	3.739713
the text answer	5.000000	3.541465
everywhere in the	5.000000	4.264479
in a group	5.000000	5.055714
circle the animals	5.000000	3.500743
can be found	5.000000	3.935567
do not have	5.000000	5.247205
types of insect	5.000000	6.014156
text answer the	5.000000	3.541465
are known for	5.000000	3.846374
vertebrates have backbones	5.000000	7.200905
is an animal	5.000000	3.955639
use the words	5.000000	3.956662
most of the	4.000000	4.954129
active at night	4.000000	5.038278
vertebrates vertebrates have	4.000000	7.874008
their lives in	4.000000	4.742366
they have a	4.000000	4.689067
a backbone and	4.000000	4.553202

Type	▼ Frequency: 01 - Freq	Dispersion: 01_CV
ticks and mites	4.000000	4.884998
in the box	4.000000	4.650365
they are also	4.000000	5.778201
all of the	4.000000	4.711899
that they use	4.000000	5.712939
animals that live	4.000000	4.842426
the box to	4.000000	4.650365
live in groups	4.000000	7.874008
the world except	4.000000	4.769286
among the most	4.000000	5.300032
kind of animal	4.000000	5.963591
all birds have	4.000000	5.267944
take care of	4.000000	5.062369
not have a	4.000000	6.072326
1 what is	4.000000	4.373073
on the outside	4.000000	4.813270
the words in	4.000000	4.650365
closely related to	4.000000	4.919070
they are related	4.000000	4.313670
they live in	4.000000	5.735603
this means that	4.000000	5.571655
are the only	4.000000	5.343576
of their lives	4.000000	4.742366
the largest living	4.000000	5.697136
lungs live on	4.000000	7.874008
are related to	4.000000	4.313670
they do not	4.000000	4.192743
words in the	4.000000	4.650365
help each other	4.000000	7.874008
to keep them	4.000000	4.243512
these animals are	4.000000	4.731002
are closely related	4.000000	4.139159
look at the	4.000000	4.858444
today there are	4.000000	5.697098
the same animal	3.000000	4.822382
over the world	3.000000	6.141360
million known species	3.000000	5.987682
they are the	3.000000	6.473027
found almost everywhere	3.000000	6.599412
75 percent of	3.000000	5.987682
2 what is	3.000000	5.120161
is in danger	3.000000	5.437691
the name of	3.000000	5.448135
like their parents	3.000000	5.523216
all animals are	3.000000	5.987682
it is also	3.000000	4.579777
though they cannot	3.000000	6.661270
for more than	3.000000	6.191844
scientists are constantly	3.000000	5.987682
silverfish and bees	3.000000	5.987682
draw your favorite	3.000000	6.057243
webs to catch	3.000000	4.563359

6 4-grams

Type	▼ Frequency: 01 - Freq	Dispersion: 01_CV
is in danger of	3.000000	5.443136
lay eggs lungs live	3.000000	7.874008
lived on earth for	3.000000	6.201262
insects developed on earth	3.000000	6.004001
look at the pictures	3.000000	5.219482
flies grasshoppers silverfish and	3.000000	6.004001
are among the most	3.000000	6.093550
teeth that they use	3.000000	5.540263
all animals are insects	3.000000	6.004001
animals in fact about	3.000000	6.004001
do not have a	3.000000	5.652698
silverfish and bees are	3.000000	6.004001
are about 1 million	3.000000	6.004001
butterflies beetles ants flies	3.000000	6.004001
insects insects developed on	3.000000	6.004001
live on land and	3.000000	6.042638
eggs lungs live on	3.000000	7.874008
a species or type	3.000000	6.653324
to help each other	3.000000	7.874008
largest group of animals	3.000000	6.004001
fact about 75 percent	3.000000	6.004001
animals are insects insects	3.000000	6.004001
before humans did today	3.000000	6.004001
the largest group of	3.000000	6.004001
75 percent of all	3.000000	6.004001
bees are all insects	3.000000	6.004001
the insects are the	3.000000	6.004001
species or type of	3.000000	6.653324
have lived on earth	3.000000	6.201262
of insect and scientists	3.000000	6.004001
discovering new species butterflies	3.000000	6.004001
scientists are constantly discovering	3.000000	6.004001
species butterflies beetles ants	3.000000	6.004001
known species or types	3.000000	6.004001
mammals that live in	3.000000	7.874008
constantly discovering new species	3.000000	6.004001
did today there are	3.000000	6.004001
almost everywhere in the	3.000000	6.601114
and bees are all	3.000000	6.004001
not have a backbone	3.000000	6.372989
or types of insect	3.000000	6.004001
1 million known species	3.000000	6.004001
on earth for more	3.000000	6.201262
spend most of their	3.000000	5.530194
of all animals are	3.000000	6.004001
are closely related to	3.000000	4.769630
humans did today there	3.000000	6.004001
beetles ants flies grasshoppers	3.000000	6.004001
and scientists are constantly	3.000000	6.004001
is the name of	3.000000	5.455544
million known species or	3.000000	6.004001
developed on earth long	3.000000	6.004001

Type	▼ Frequency: 01 - Freq	Dispersion: 01_CV
group of animals in	3.000000	6.004001
about 75 percent of	3.000000	6.004001
about 1 million known	3.000000	6.004001
their lives in water	2.000000	6.820170
as large as a	2.000000	5.555711
a backbone and they	2.000000	5.534838
that are found all	2.000000	6.783953
help each other care	2.000000	7.874008
are related kinds of	2.000000	7.874008
spiders are eight-legged creatures	2.000000	5.670645
draw a line between	2.000000	6.351555
for making silk webs	2.000000	5.670645
what kind of animal	2.000000	7.874008
part of their lives	2.000000	5.600640
all have wings though	2.000000	7.874008
that is in danger	2.000000	5.562013
say its name out	2.000000	5.531810
on outside sources to	2.000000	5.534838
to the group of	2.000000	6.058216
that can produce milk	2.000000	7.874008
poisonous snakes that have	2.000000	7.874008
that spend most of	2.000000	5.752634
of flying insects the	2.000000	7.874008
mammals that belong to	2.000000	7.874008
grows hair at some	2.000000	7.874008
pigeons in big cities	2.000000	7.874008
some animals eat only	2.000000	7.874008
long periods of time	2.000000	7.874008
have glands that can	2.000000	7.874008
to the animals that	2.000000	5.523025
are small mammals that	2.000000	5.614016
animals that have feathers	2.000000	6.591190
what is a baby	2.000000	7.874008
live in groups to	2.000000	7.874008
have dry skin covered	2.000000	6.704868
scientific name of the	2.000000	7.874008
animals live in groups	2.000000	7.874008
fawn mouse calf chick	2.000000	6.131854
moths are related kinds	2.000000	7.874008
animal that breathes air	2.000000	7.874008
scorpions ticks and mites	2.000000	6.327378
in a variety of	2.000000	7.874008
there are about 38,000	2.000000	5.670645
silk webs to catch	2.000000	5.670645
are invertebrates that live	2.000000	7.874008
answer the questions 1	2.000000	5.593328
periods of time without	2.000000	7.874008
answer the questions how	2.000000	5.533614
animals stand on the	2.000000	7.874008
box to complete the	2.000000	5.553284
birds are the only	2.000000	7.874008
the most intelligent of	2.000000	7.874008

Source: #LancsBox 6.0.

Appendix P Learners' samples of language production: *Neighborhood* (T1)

Observation: the numbers on the left refer to the number a student has in each of the classes` class roll call.

Individual learner production in 4A and 4B groups

4A	Learner samples - class and homework
403	There are different types of medicines.
	This is because we use different medicines for different reasons
404	There is a restaurant nearby
407	Find the best road to the police station
414	City block with park and mall
424	Find the best road to your apartment

4A	Learner samples - posttest
401	Find your way around a supermarket
	There is a park (mall, supermarket, school, park) nearby
402	Find the best road to
403	City block with apartments and buildings
406	There is a park nearby
	Find the best road to your ...
409	Find your way around a park
414	A school map, for example, help you plan
	There is a park nearby
415	Tall house where many people live
	They may be able to walk..
419	Find the best road to your park
421	There is a hospital nearby
422	City block with homes and stores

4B	Learner samples - class and homework
432	They may have parks, malls and shops
433	There is a Mall nearby
435	They may have a high school
438	They might need a house with more space
440	There is a shop nearby
444	There are many different types of shops
445	There is a pet shop nearby

4B	Learner samples - posttest
427	Find the best road to your friend`s house
	There is a library nearby
428	There is a hospital nearby
429	Find the best road to your pet shop
	City block with ...
432	Find your way around a pet shop
	Find the best road to your Mall
	City block with houses and shops
435	There is a bakery (hospital, park, bank) nearby
447	Find your way around your neighborhood

Source: Learner data from tests and classwork.

Individual learner sentences in 5A and 5B groups

5A	Learner samples - class and home work
501	There is a shop nearby (park, pet,shop, house, supermarket) nearby
503	Find the best road to parks
504	There is a bank nearby
506	City blocks with ...
513	The Mall where many people go
516	City block with church and bank
517	There are many different types of houses (parks, pet shops, library)
518	City blocks with ...
519	A Library is a wonderful place to visit
524	Tall houses where many people live

5A	Learner samples - posttests
502	Tall apartment where many people live
505	A Shopping is a wonderful place to visit
510	There is a park nearby
511	Find your way around the library
514	They may be able to walk to scholl and pet shop
523	They may also use public transportation to get to different parts of the city
524	There is a bakery nearby

5B	Learner samples - class and home work
526	Do you have friends in your neighborhood?
527	City block with pet shop and stores
529	Do you have high school in your neighborhood?
531	There are many different types of restaurants
547	They may be ablr to walk to school, the house, th pet shop, the park, ...
548	There is a Mall nearby (a park, a restaurant, a pet shop)

5B	Learner samples - posttest
526	Find your way around a restaurant
527	City block with park and restaurant
531	Tall apartment buildings where many people live
533	Do you have friends in your neighborhood?
537	Find the best road to your friend`s school
	Tall houses where many people live
	There is a park nearby
540	Find your way around a school
544	There are many different houses (shops, ice cream shops)
547	Find the best road to your school

Source: Learner data from tests and classwork.

Individual learner sentences in 6A and 6B groups

6A	Learner samples - class and home work
612	They may be able to walk to shop. The Mall and Park
	Who help kids to get to school?
	There is a shop nearby
613	When there is a picnic at the school, you clean up after yourself
	A ... map, for example, helps yo plan
616	The bakery is a wonderful place to visit
	There are many different types of ..... (supermarkets, houses, schools,

6A	Learner samples - posttest
612	Find your way around a nature park
	There is a bakery nearby
614	Find the best road to your Mall
616	Find the best road to your house
	Do you have frineds in your city?
621	They may be able to walk to the farm (the bank, the park and mall)

6B -	Learner samples - class and home work
630	They might need a pet shop
632	Find your way around a house
633	Do you have friends in your school?
	They may be able to walk to the library
634	If you were going on a hike in a zoo, you would need a map
635	They may be able to walk to school (to the park, to the house)
	(to the store, to the restaurant, to the office, to the bakery)
	The park is a wonderful place to visit
636	There is a supermarket nearby

6B	Learner samples - posttest
632	They may also use bus to get to different part of the city
	City block with homes and a library
	There is a restaurant nearby
633	Find the best road to your family`s buildings
638	Find your way around a pet shop
640	Tall achool buildings where many people study
644	When you have a picnic at the school you clean up after yourself
646	Tall house where many people live
	Find your way around a hospital
	Find the best road to your friend`s apartment

Source: Learner data from tests and classwork.

Appendix Q Learners' samples of language production: *Animals* (T2)

Observation: the numbers on the left refer to the number a student has in each of the classes` class roll call.

Individual learner sentences in groups 4A and 4B

4A	Learner samples - class and homework
404	The giraffe eats a lot of leaves from the trees with its big neck
	The macaw moves by flying
405	I like gold fish
408	The horse run much fast
409	Tiger hav legs, long tail and run
411	I don't have dog
	Lion is a good hunter
412	The lion is the king
	Birds have feathers, fly
414	Some monkeys are small
415	Dogs have different breeds
	Giraffe have a long neck, walk
419	The lion is the king of the jungle
424	The tiger likes to eat meat
425	A dog is a very smart animal

4A	Learner samples - posttest
401	Gazela are animals that have strong legs to run
404	Rabbits are jumping animals
	Turtle are kind of animal that lives in water
405	Lions are animals that have strong legs to run
406	Frogs are jumping animals
408	Crickets are small animals that can jump very
414	Spiders are insects that are found almost everywhere in the world
415	Frogs are jumping insects
419	Frogs are jumping animals
421	Butterflies, parrot and bees can fly
	Lions are animals that have strong legs to run
425	Many people have dogs as pets

4B	Learner samples - class and home work
432	The tiger is a fast animal that **moves running**
	The bird is a beutiful animal that **moves flying**
433	The dog has four paws, tail, fur.
	The fish has scales, gills. Flippers
434	The cat is smart
436	The pig like to play in the mud
	The duck love water
	Cat is furry with four legs. It walks and run
439	The birds are beautiful and colorful
440	The lion has a beautiful mane
441	The giraffe's neck is very long
	The elephants are very big and strong
	The lion's mane is so fluffy
443	When I went to the zoo I saw na elephant
447	The elephan has big ears and a trunk
449	The bird is colorful and flies
	The fish is golden and swims
	The cat is fluffy and walks
450	Lion has a big mane and he is very fast
	Fish has scales, hs flippers, to help swimming and moves swimming
451	The ducks like swimming

4B	Learner samples - posttest
427	Shark (fish, octopus) are **a kind of animals that live in the water**
429	Crickets, ants and bees **are all insects**. They have 6 legs.
430	Zebras a**re animals that have strong legs to run**
432	Shark are **a kind of animal that lives in the water**
433	Lion are **animals that have strong legs to run**
435	Giraffe are animals that have strong legs to run
436	Birds are cute animals
439	Frogs and rabbits are **jump**
440	Lion are animals that have strong legs to run
441	Rabbits are medium **animals that can jump very well**
442	Parrots are animals **that have feathers. They lay eggs.**
447	The zebra is striped in black and white
450	Hippos live in parts of Africa
451	Shark are a kind of animal that lives in the water

Source: Learner data from tests and classwork.

Individual learner sentences in 5A and 5B groups

5A	Learner samples - class and homework
501	Horses have hooves and are fast
	Elephant have a trunk
502	I admire the beauty of the giraffe
508	Monkey love banana and jump
509	A fish has gills and scales
	An octopus has tentacles
512	My dog is very messy and loves to walk
514	Lions are wild animals and they live in the jungle
520	The dog is a cute, affectionate and smart animal
524	Crickets are jumping insects
	Insects developed on Earth before human did

5A	Learner samples - posttest
502	Shark are **a kind of animal that lives in water**
504	Elephants **cannot jump**
509	Shark are **a kind of animal that lives in water**
511	Kangaroos are big animals can jump vey well
512	Parrot are **birds that cannot** swim
513	Sharks live in water
518	Chickens are birds that cannot fly
519	Blue parrots live in parts of Brasil
520	Dogs are animals that are found almost everywhere in the world
521	Some insects can jump

5B	Learner samples - class and homework
526	The fish **swim**s and has a very beautiful fin
	The dog he walks **has a tail and 4 paws**
527	The giraffe's **neck** is too big
530	The lion **has a mane**
536	I love cats but I don't have one
541	The cat **can climb and can walk and run**
	The tiger is **carnivorous and can run**
542	The cat love to sleep

5B	Learners samples - posttest
525	Horses are animals that have strong legs to run
	Shark are a kind of animal that lives in the water
526	Starfish are a kind of animal that lives in the water
527	Rabbits ae jumping animals
528	Horses are a kind of animal that lives in the farm
529	Shark are a kind of animal that lives in the water
530	Rabbits are jumping animal
532	Rabbit are small animals that can jump very well
536	Frogs are jumping insects
537	Rabbits (dog, cat) are a kind of animal that lives in the house
538	Shark are a kind of animal that lives in the water
539	Lion a kind of animal that lives in the zoo
541	Penguins are birds that cannot sing
546	Some birds cannot fly

Source: Learner data from tests and classwork.

Individual learner sentences in 6A and 6B groups

6A	Learner samples - class and home work
601	Shark live in the ocean
	The bird isn't beautiful
603	The pig likes mud
	The monkey loves bananas
603	The butterfly has wings of many colors and it fly
606	The dog besides being my favorite animal is also one of the most beautiful
	The elephant is one of the largest animals
	in the world ever known and is also the strongest
608	I have an allergy to cats
609	The cat eat fish
610	Giraffes live in parts of Africa
615	The lion normally eats meat
623	The tiger have a whiskers and spots

6A	Learner samples - posttest
601	Camel live in parts of Egito
603	Rabbits are small animals that can jump very well
604	Tigers are animals tha have strong bite
606	Many people have dogs as pets
611	Shark are a kind of animal that lives in the water
613	Ducks are animals that have feet. They lay eggs.
615	Many people have rabbits as pets
620	Lion is the largest of the cats
621	Giraffes live in parts of Africa

6B	Learner samples - class and homework
624	Cat walk and run, dog walk and run, bird fly, duck swim
625	The dog has paw
	The lion has a tail and a mane
	The fish has gills
	The tiger moves running
	Butterfly moves flying
627	Fish live in the ocean
	The horse eat plants
	The monkey eat bananas
628	The duck swim, the bird fly, the tiger run
	The turtle swim and walk
632	Parrot fly, not run
	Shark swim
633	The dog barks
	The horse run very fast
	The bird flies and is cute
	The fish swim
634	The cow has spots
	The giraffe walk and run
636	The elephant has a trunk and is very big
	The bird has a beak and is very colored
637	The mane of the lion is beautiful
638	The dog bark and bites
	I hate pigs
639	The cat has long whiskers
	The tiger is stripes
	The dog runs very fast
	The pig love to run in the mud
	The bird fly in the sky
644	A bird fly with feathers
	The lion have a mane
646	Ca'ts meows are much finer than dog's bark

6B	Learner samples - posttest
624	Rabbit are small animals **that can jump very well.**
	Horse are animals that **have strong legs to run**
625	Chickens are **birds that lay eggs**
627	The butterflies have two wings
629	Sharks live in the sea
630	Lion is the forest king
631	The shark eat the octopus
632	Chickens are birds **that cannot fly**
633	Shark bites
	Rabbit can **jump very well**
635	Kangaroos **jump** high
641	Shark lives in the water
642	Fish live in the water and can be yellow, blue, red
643	Aligators live in the swamp