# Exploring the Content for Content and Language Integrated Learning (CLIL) Programs for Social Sciences and Humanities

*M.D.S.S. Kumara, D.A. Wehella, H.P.K. Pathirana, R.M.L.C. Kumari, A.M.C.K Abeysekara, T.S. Abeynayaka, and T.S. Rajapaksha*

*Department of English Language Teaching, Rajarata University of Sri Lanka, Mihintale.* deltcentrearusl@gmail.com

## 1. Introduction

Content and Language Integrated Learning-CLIL (Coyle et al., 2010) is a method of teaching a language using the content of various subjects taught in educational institutions. It is an integrated approach to language learning based on the premise that language must be used as a medium for learning content, and content as a resource for language learning. In this method of instruction, attention is shifted from learning a language by itself to learning a language through a relevant learning context. Thus, communicative competence is achieved in the act of learning specific topics such as Science, Economics, Management etc.

CLIL is based on the assumption that learners learn a second language best when they are given a language in a meaningful, contextualized form with the primary focus on acquiring information. Since English Language Teaching (ELT) programs in Sri Lankan universities primarily focus on preparing learners for academic studies, CLIL programs are ideal for this purpose. The best CLIL programs require a close collaboration between the language teacher and the content teacher, who are experts in one's own field. The language teacher's responsibility in a CLIL program is to teach the target language, support the content teacher through the introduction of vocabulary and functional language relevant to the target subject, and encourage critical thinking (cf. Bridge Education, 2002-2023).

The majority of the ELT programs conducted at the faculties of Agriculture, Applied Sciences, Management Studies, Medicine and Allied Sciences, and Technology of the Rajarata University of Sri Lanka, of which the primary medium of instruction is English, show features of CLIL programs. In contrast, the ELT programs conducted at the Faculty of Social Sciences and Humanities of the Rajarata University of Sri Lanka (FSSH-RUSL), of which the primary medium of instruction is Sinhala, focus on achieving everyday English communication skills; i.e., they are General English programs. However, the University Grants Commission of Sri Lanka has recently directed the faculties of Social Sciences and Humanities to start offering their degree programs in the English medium. In the process of offering the degree programs of FSSH-RUSL in the English medium, the ELT teachers can support the content teachers by introducing vocabulary and functional language relevant to the content subjects.

The present study was conducted with the overall aim of examining the most frequent English vocabulary and functional language relevant to the subjects offered by FSSH-RUSL by compiling and analyzing a digitalized corpus of Academic English: Rajarata University Social-sciences and Humanities (RUSH) corpus. The specific objectives of the study are to recognize English medium written texts that represent the subjects offered by the 08 departments of FSSH-RUSL, to compile a corpus of written academic discourse by digitalizing representative samples of the texts recognized thus, to make a linguistic analysis of the compiled corpus in order to recognize the vocabulary and functional language that could be used in CLIL teaching materials for the ELT programs of FSSH-RUSL, and to make suggestions for the improvement of the ELT programs of the faculty based on the findings of the study. What is meant by a 'corpus' here is a collection of authentic language data collected for linguistic study, and stored and accessed electronically. Software used to analyze corpus data are termed 'concordancers'. Corpus Linguistic methods use both quantitative and qualitative data analysis.

## 2. Materials and Methods

Since the key premises in CLIL programs are collaboration and integration, the first step in the present study, i.e., the selection of written academic texts representing the subjects offered by the 08 departments of FSSH-RUSL, was done in liaison with the coordinators of content subjects. Thus, written texts of the text type Printed Books (PB) were selected from the following content disciplines amounting to a generalized word count of 50,000 from each, totaling a raw word count of 645,511 (MS-word count) to compile the version 1.0 of RUSH corpus with the file name convention - Corpus name+Text Type+Subject Code: Archeology (RPBAR), Economics (RPBEC), Education (RPBED), Environment Management (RPBEM), History (RPBHI), Information Technology (RPBIT), Languages and Linguistics (RPBLL), Management (RPBMA), Mass Communication (RPBMC), Sociology (RPBSO), Statistics (RPBST), Tourism (RPBTO), and Water Resources Management (RPBWR).

The selected texts were then digitalized by scanning them, uploading them to a google drive (Google LLC, 2020), and opening them using the Google-Docs tool of the Google Drive. The digitalized data were saved onto the computer first as .docx files, and then they were converted and saved as concordance-readable .txt files to compile the sub-corpora of the RUSH corpus. Metadata for each sub-corpus file including socio-biological information of the authors were separately saved. The analysis of the RUSH corpus data was carried out using the online concordancer- Lextutor (Cobb, 2022), and the open-source concordancer- AntConc 4.3.1. and 3.4.1 (Anthony, 2024). Among the corpus analysis tools which can be used to explore content vocabulary and functional language, Keywords, Concordance lines or Key Words in Context (KWIC), Collocates, Frequency Range, and Lexical bundles (N-grams) are prominent.

## 3. Results and Discussion

Keywords are lexical items that are far more frequent in a (specialised) corpus compared to a reference (general) corpus. They indicate topic specific vocabulary and grammatical items that will reveal more specific information about the language preferences of the particular discipline. Thus, keywords are highly useful in preparing ELT material in CLIL programs. Table 1 below shows the first 20 keywords of the RUSH 1.0 corpus obtained by uploading the entire corpus file to Lextutor. The 'Keyness Factor' given in the table is the number of times more frequent the relevant word is in the present corpus than it is in the reference corpus used by Lextutor (bnc_coca_fams_speechwrite_US_UK_per10mill). For example, the first item in the output **9827.00 meaning** means that **meaning** has **1** natural occurrence in 10,000,000 words reference corpus, but **592** occurrences in the RUSH 1.0 (602,393-word) text -- or, (592/602393) x 10,000,000 = **9,827** occurrences if the present text were the same size as the reference corpus. The word is thus 9,827 / 1 = **9827.00 times** more frequent in the present text than it is in the reference corpus. This probably means the word plays an important (or 'key') role in the present text (cf. (Cobb, 2022). Keywords of all 13 sub-corpora of the RUSH 1.0 corpus were separately extracted this way, but the page limit restriction of the present paper doesn't permit reporting them.

*Table 1. The first 20 keywords of the RUSH 1.0 corpus extracted from Lextutor.*

| S. NO. | KEYNESS | KEYWORD | S. NO. | KEYNESS | KEYWORD |
|---|---|---|---|---|---|
| 01 | 9827.00 | meaning | 11 | 1428.00 | eventual |
| 02 | 9628.00 | relation | 12 | 1361.00 | motivate |
| 03 | 8300.00 | politic | 13 | 1179.00 | sinhalese |
| 04 | 7387.00 | compute | 14 | 1129.00 | anthropogenic |
| 05 | 5777.00 | program | 15 | 1129.00 | archaeological |
| 06 | 4648.00 | irrigate | 16 | 1062.00 | professional |
| 07 | 3287.00 | especial | 17 | 996.00 | papyrus |
| 08 | 2473.00 | situate | 18 | 996.00 | moghul |
| 09 | 2291.00 | probable | 19 | 930.00 | criterion |
| 10 | 1992.00 | equip | 20 | 913.00 | excavate |

In order to know the context and regular patterns of the usage of the above keywords, the basic tool of the software called 'concordancing', which shows "(Lexico)-Grammatical Co-occurence" (Gries, 2009), can be used. This list of lines with the search word at the centre is also called 'key word in context' (KWIC), and is useful for exploring different meanings of the search word. Figure 1 below shows the KWIC lines for the keyword of the corpus 'irrigate', occuring in Water Resources Management (RPBWR) sub-corpus as extracted by AntConc 4.3.1. sorted by Right collocates. As depicted in the figure, different forms of the verb 'irrigate' are presented here with the words frequently occurring with the word (collocates) so that both the ELT teacher and the learner find it easy to acquire the meaning of the word.

| | File | Left Context | Hit | Right Context |
|---|---|---|---|---|
| 1 | RPBWR.txt | stems and more diversified irrigated agriculture. The history of | irrigated | agriculture in Sri Lanka has essentially been one of |
| 2 | RPBWR.txt | command. The scheme would allow intensification of existing | irrigated | agriculture on approximately 9,100 net hectares and developm |
| 3 | RPBWR.txt | &M under modernised irrigation systems and more diversified | irrigated | agriculture. The history of irrigated agriculture in Sri Lanka |
| 4 | RPBWR.txt | collection in Sri Lanka under traditional systems of rice based | irrigated | agriculture. The second is concerned with O&M under |
| 5 | RPBWR.txt | and (f) the component small tanks as well as the | irrigated | rice lands are shown in the same figure. Itakura |
| 6 | RPBWR.txt | MW. About 10,000 acres of the total area constituted existing | irrigated | rice lands. About 25,000 acres of the land was situated |
| 7 | RPBWR.txt | and right bank) and a system of distributory canals, which | irrigated | the entire area below their command. Thus, both well- |
| 8 | RPBWR.txt | mand. Thus, both well-drained and poorly drained lands were | irrigated. | The momentum was stepped up after Independence through a |
| 9 | RPBWR.txt | ncement of the ancient "hydraulic civilisation". The absence of | irrigated | upland agriculture practiced on a sustained basis over this |
| 10 | RPBWR.txt | been tried on an experimental basis, including an attempt to | irrigate | upland areas with lift irrigation. Another was the use |
| 11 | RPBWR.txt | ster Plan, which proposed to construct fifteen reservoirs and to | irrigate 900,000 | acres over a period of 30 years, was Rs. 5.583 billion |
| 12 | RPBWR.txt | two main canals. The Left Bank canal was designed to | irrigate | an area of 80,000 acres while the Right Bank canal |
| 13 | RPBWR.txt | evels in 1980 and prospective levels with planned expansion in | irrigated | area during the 1990s. It reviews past strategies for |
| 14 | RPBWR.txt | potential new irrigable area is located uphill of the existing | irrigated | area. The proposed irrigation development would include: - im |
| 15 | RPBWR.txt | above the village. That tank, however, was not used to | irrigate | land. On the contrary, its express purpose was to |
| 16 | RPBWR.txt | the fourth order streams there is sufficient discharge from the | irrigated | lands during the February-April period that would ensure |
| 17 | RPBWR.txt | state sector. The latter predominate and are synonymous with | irrigated | settlement schemes. Gravity irrigation schemes can be classifie |
| 18 | RPBWR.txt | opment Board (RVDB). The prime objective was to develop and | irrigate | the 30,00 acres for double cropping, enhance yields, accommo |

*Figure 1. KWIC lines for 'irrigate' in Water Resources Management sub-corpus as extracted by AntConc 4.3.1.*

Another common tool in corpus analysis software is 'word lists' or 'frequency lists', which usually provides the lists of words in the corpus according to their frequency. With Lextutor's 'Range' tool, it is also possible to extract the distribution range of the words across the sub-corpora of the corpus. Using Lextutor's 'Stoplists' function, the distribution range of discipline-specific vocabulary items, which is beneficial in the curriculum design of the CLIL programs could also be obtained. Table 2 below depicts the distribution range of the top 10 discipline-specific words of the three sub-corpora of the RUSH 1.0 corpus representing the three disciplines under the Department of Humanities of FSSH-RUSL, viz., Education (T1), History (T2), and Tourism (T3) as extracted by Lextutor. The first column gives the word family (Fams); the second column-the frequency in all three sub-corpora; the third- the distribution range; the fourth-the vocab profile (VP) group number; and the last three columns give the subcorpora in which the word appears, with the frequency given within brackets. The ELT teacher in a CLIL program for the Department can make crucial decisions regarding the lexis for a common ELT program based on this kind of results.

A corpus linguistic tool that can be used to explore functional language useful in a CLIL program is 'Lexical Bundles', which are also termed 'Gram Bundles' or 'N-grams'. They are recurring word strings that provide insights into the way words in a corpus go into combinations with other words frequently. For example, the top 10 most frequent 4-word N-grams of the RUSH 1.0 corpus, extracted by AntConc 3.4.1, given in Table 3 below show that the most frequent lexical bundles in the corpus are prepositional phrases which constitute an important type of functional language in an ELT program.

*Table 2. The distribution range of the top 10 discipline-specific words of Education, History and Tourism sub-corpora of RUSH 1.0 corpus as extracted by Lextutor.*

| Word family | FREQUENCY | RANGE | T1 | T2 | T3 |
|---|---|---|---|---|---|
| curriculum | 249 | 2 | T1(248) | | T3(1) |
| philosophy | 171 | 3 | T1(147) | T2(17) | T3(7) |
| critic | 163 | 3 | T1(146) | T2(11) | T3(6) |
| concept | 152 | 3 | T1(68) | T2(13) | T3(71) |
| analyse | 144 | 3 | T1(83) | T2(5) | T3(56) |
| dynasty | 122 | 1 | | T2(122) | |
| sustain | 122 | 3 | T1(8) | T2(12) | T3(102) |
| theory | 111 | 3 | T1(79) | T2(9) | T3(23) |
| ancient | 110 | 2 | | T2(101) | T3(9) |
| define | 109 | 3 | T1(52) | T2(26) | T3(31) |

*Table 3. Top 10 most frequent 4-gram bundles of RUSH 1.0 corpus as extracted by AntConc 3.4.1.*

| Rank | FREQUENCY | N-GRAM |
|---|---|---|
| 1 | 88 | the end of the |
| 2 | 86 | at the same time |
| 3 | 74 | in the united states |
| 4 | 69 | on the other hans |
| 5 | 60 | history of mass communication |
| 6 | 59 | a history of mass |
| 7 | 55 | The company x s |
| 8 | 52 | as a result of |
| 9 | 52 | in the case of |
| 10 | 48 | in the form of |

## 4. Conclusion

In fulfilling the responsibility of an ELT teacher of a CLIL program to support the content teacher through the introduction of discipline-specific vocabulary and functional language, corpus linguistic tools provide a reliable means of empirically recognizing the required language. The balanced samples of language included in the corpus, and the mixed method analysis provided through the tools like Keywords, KWIC, Collocates, Frequency Range, and N-grams, made it obvious that the results of the present study help both the ELT teacher and the content teacher perform their function easier than attempting to recognize the required language by manual means. Continuous collaboration of both teachers in the implementation of the CLIL program is, nevertheless, crucial in order to make the teaching program totally effective.

## 5. Keywords

CLIL, Concordance, Corpus, Vocabulary

## 6. References


- Anthony, L. (2024). AntConc (Version 4.3.1, and 3.4.1) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software
- Bridge Education, (2002-2023). What Is CLIL? The Global Trend in Bilingual Education Explained
- [Blog post]. https://bridge.edu/tefl/blog/what-is-clil/
- Cobb, T. (2022). *Compleat Lexical Tutor* [computer program] https://www.lextutor.ca/
- Coyle, D., Hood, P., & Marsch, D. (2010). *Content and Language Integrated Learning*. Cambridge University Press.
- Google LLC . (2020). Google Drive. https://www.google.com/drive/
- Gries, S. T. (2009). *Quantitative corpus linguistics with R : a practical introduction*. Routledge.