Corpus Linguistics And Technology Integration In SLA; Generating Language Tasks Through SKETCH Engine

Dr. Juliana Çyfeku

Faculty of Education and Philology Foreign Language Department University of Korcë, Albania
juliana.cyfeku@gmail.com

Abstract

Given that, Corpus Linguistics supports the integration and the use of sociocultural approaches in Second Language Acquistion (SLA), this study underpins what corpus linguists uphold in their researches concerning language and its features to support and assist language teachers and learners in the acquisition of the English. The main focus of this study goes in line with consistent SLA theories, particularly those that emphasize the sociocultural approaches targeting: input and output (linguistic data), use of authentic texts, the importance and creation of real-world language tasks, learner-learner interaction, and learner autonomy in dependency to explore the data provided in corpora contexts.

In relevance to technology integration in SLA, SkELL (Sketch Engine for Language Learning) is a free, open online search engine that shows how genuine language as selected for linguistic purposes, is used by real speakers of the targeted language. The present study aims to highlight the practicality of implementing: Word Sketch, Thesaurus, Wordlist as innovative tools among a range of others. It addresses the use of relia and authentic texts that guide autonomously language learners undertake, explore and reach to conclusive remarks to any relevant issue-specific empirical study of the language.

Keywords: Corpus Linguistics, Technology Integration, Sketch Engine, Word Sketch, Thesaurus, Wordlist

1. INTRODUCTION

The present study is written primarily for English teachers, English Language Teaching students (ELT) at university level, to those who have taken a CL course in teacher preparation as it is the case of ELT master's programme students where I teach the course "Corpora in Applied Linguistics" at the Foreign Languages Department, to those practitioners who may have a limited background in Corpus Linguistics (CL), or to those researchers/language learners who are unwilling to implement in the SLA being resistant to the fact that they may not be skilful enough to integrate technology as an integral part of today's educational trend in reference foreign language learning.

Nowadays, language corpora are widely being used by all kinds of language users including: teachers, students, researchers, curriculum designers and ELT (English Language Teaching) / EAP (English for Academic Purposes) / ESP (English for Specific Purposes) learners. Computers and technology have become widely available and vitally serve the scope of being informed to all of the above language practitioners in homes, schools or any virtual language learning area. Due to the mutual

connection that these two novice and up-to-date means of communication compliment one-another especially in foreign language acquisition, affect and require any possible development to be disseminated in what has made use of Corpus Linguistics. Having the attributes of outstanding sociocultural approaches, Corpus Linguistics has its main focus on the "naturally occurring texts" (Reppen, 2010). Specifically, in the world of Corpus linguistics, a corpus is a large, principled collection of naturally texts (written or spoken) stored electronically. The texts provided reflect the nature of genuine discourse obtained from actual language situations, such as people conversing or chatting either formally or informally in classroom settings, class assignments, meetings, telephone conversations, essays, speeches, magazines, newspapers and/or books. Within the framework of corpora authenticity, principal to the statement that any online and purchased type should be guided from all kind of corpus users and repeatedly make decisions about the language occurrences; eliciting among a range of texts which linguistic features and what kind of field-specific vocabulary to teach and /or to test. In recent years, with the advent of computers and the galloping speed of implementation of technology especially in educational settings, EFL/ ESP / EAP professionals have embraced and differentiated the use of "authentic" materials, presenting language from natural texts rather than made-up examples (McDonough & Shaw, 1993).

On purpose the present study reflects and supports what Geoffrey Leech (1997), and Römer (2008) uphold their remarks on the significance of indirect and direct applications of approaches of CL in the field of foreign language teaching and learning. Indirect and direct approaches to corpora provide insights of the nature of language as displayed in corpus-based texts and as such informing about language use in various ways

learners, teachers, researchers and language material designers. Direct approaches guide these language users on authentic, hands-on activities to generate tasks for novel classroom material. In her 2001, book Tonini-Bonelli, explicates that the two approaches corpus-based or corpus-driven posit research by providing clues to critical observatory of the language applications. An alternation of the practical tasks generated from the corpora tools can reveal unprecedented insights to etiquette language as a system of patterns with enormous examples of word combinations. The latter are furnished beyond the learner's perception with features of inquisitive grammar intertwined with the lexical items in any given foreign language. These classroom or digital implications from any available learning environment in line with technological appliances that every individual daily possesses, can be instructed by the teacher or opted to provide space to any interested parties to suit corpora language to their needs and linguistic expectations as well.

Indirect approaches to using corpora in the language SLA have made use from linguistic research by drawing comparisons from the nature of language provided in various course books and the 'naturally occurring language' (Reppen, 2010) typical of corpora in written and oral discourse. From the eye of a researcher the perspective of corpus-based language relies on the frequency of occurrence of linguistic structures guiding so our issue-specific study on those problematic structures extracted straight from a native speaker's mouth reflecting real-life language use. For instance the frequency rated for the commonly used language patterns in language textbooks may be searched in any chosen corpora. The obtained data from the sample contexts aims to provide certain contextual clues for the most frequently used language patterns guiding the inquiry independently to delve and reach to assumptions of agreements or

disagreements extracted as excerpts from corpus data (Oztekin, E. & Candan, E. 2019).

2. METHODOLOGY

This study provides some practical activities assuming that you should have access to a reasonably large corpus such as: BNC, BAWE, COCA, Academic Abstract, Academic General, MICASE, Time Magazine Corpus,..etc

Walsh (2014) provides a list of "Engaging Uses of Instructional Technology" posted in EmergingEdTech that highlight the contribution of CL tools in successful language learning. Essentially, instructional technology aims at targeting the active reactions and subsequent interaction from the learner's part towards data-driven language learning. Thus, turning language learners into autonomous researchers who will embrace language discovery (Johns, 1986; Barlow, 1996; Tribble & Johns, 1997). In addition, what Pérez-Parades (2019) insinuates is that methods of research in Corpus Linguistics can be implemented into the language classroom to foster the independent acquisitions of language constituents (lexis, grammatical constructions, collocations, clusters,, and so on). Being entirely and inductively guided the authentic data governs the learners and the whole learning process to discover patterns of language that might sound unfamiliar prior to corpora use. As O'Sullivan (2007) and Lee et., al (2019) state such new, thrilling endeavour has a major impetus in fostering more complex cognitive processes in SLA acquisition that consists of making inferences and forming language-specific hypotheses and consequently enhancing language learning strategies. In relevance to data-driven language learning, psychologist Jerome Bruner (cited in O'Keeffe), coined the term 'scaffolding' to refer to the process of mediation in the learning process within a sociocultural theory (SCT) paradigm. Some key concepts of this theory centralize the learner's empowerment in taking control and self-regulating learning rather than presenting oneself as a passive transmitter of knowledge when compared to the traditional teacher prevailing language classes mostly throughout instruction. We believe in what O'Keeffe et al., (2007) argue that: […learners can be trained to operate independently to develop skills and strategies during the process to surpass instructional intervention through practice and thus become a better self-regulated learner…] (p.55).

2.1 Tools, Corpora and Online Resources

Due to the fact that both sides of the coin in the learning process: language teachers & learners, are concerned with the different contexts in observing how language can vary depending on these different contexts. By means of corpora teachers as well as learners can find valuable recourses to clarify their ambiguities or gaps of linguistic knowledge to create themselves activities and corpus-based tasks to assist the latter understand variation due to situational factors (register).

Availably, there exist a wide range of large-scale corpora and online databases that are relevant for English teachers and all kinds of learners. They may be found as ready accessible corpora or may be purchased corpora from their developers. The purpose of this study is to highlight corpora that are publicly available and moreover easily to be used by learners and implemented in our language curriculum. Friginal (2018) inferences spot out the growing interest from the corpus linguists in collecting and sharing various research groups globally, suggesting and emphasizing the merging of SLA and corpus-based research in language teaching. Beforehand it is important to note that what you need to decide prior to using which type: the research question, clear objectives and the teaching goals you require to obtain at the end of the course. Then the next step consists of the corpora

selection among a range provided on the web and the implementation of corpora tools to contemplate the targeted field-specific vocabulary and improve understanding across register variation. According to Reppen (2010, pp. 12) register is a term used to describe varieties of texts that are defined by situational characteristics (e.g., spoken vs written, edited vs. real-time). Thus the term register can be used at various level of specificity. Basically, being clear on what you intent to explore and seek responses to your questions at first you should be aware of the type of corpus you want to explore, and then decide upon the corpus tools that would easily assist your research in SLA.

Any collection of more than one text can be called a corpus, (corpus being Latin for 'body', hence a corpus is any 'body' of text. Yet the term when used in modern linguistics is well rooted in the linguistic tradition that McEnery & Wilson (2001) necessitate in the role of a corpus in a research-oriented paradigm. The principles that a good corpus ought to have are:

- ✔ representativeness,

- ✔ size,

- ✔ machine-readable form

- ✔ and standard reference.

Since linguists claim to use a corpus as a convenient way to inquire language use and social behaviour, they have to make sure to their tool i.e their language corpus and their methodology are targets of maximized representative quality of the language samples that have been included in the corpus.

2.2 Introduction to SkELL

In relevance to technology integration in SLA, SkELL (abbreviation of Sketch Engine for Language Learning,

https://skell.sketchengine.co.uk/run.cgi/skell) (Figure. 1) is a free, open online search engine that shows how genuine language as selected for linguistic purposes, is used by real speakers of the targeted language. It was first presented in 2014 when only English was supported. Nowadays, there are versions of SkELL for more than 95 additional languages. It allows language learners and teachers find authentic sentences for specific target word(s).As Karpenko-Seccobe (2021) addresses to EAP teachers, it allows learners to see lists of words which occur frequently together with a keyword (collocates). SkELL contains over 57 million sample sentences. It is very user-friendly; its creators describe it as a language-learning website 'designed to avoid scaring the students' (Kilgarriff et al., 2015, p.66 cited by Karpenko-Seccombe, 2021). SkELL searches its own corpus, which contains over 1 billion words of texts from BNC, Wikipedia and other websites, collected for the use of language learners and teachers.

When compared to other search engines on the web such as:

- ▪ AntConc, which is a freeware program that creates word frequency lists and KWICs on the web. It is an easy-to-use program that identifies n-grams of 2-6 words.

  www.antlab.sci.waseda.ac.jp/software.html

- ▪ Compleat Lexical Tutor (Cobb, 2000), which allows you freely to input texts for vocabulary analysis, a corpora and provides access to various corpora and tools consisting of many useful articles on corpora and language testing.

  www.lextutor.ca

- ▪ Web Concordancer, which is a free program that links to several corpora including Brown Corpus and Lancaster Oslo Bergen (LOB); and allows users to input and search any selected corpus.

www.edict.com.hk/concordance

- ▪ WordSmith, which is a concordancing program that, in addition to creating concoradance lines, provides other information (e.g., frequency, key words, mutual information scores, word length, etc)

www.lexically.net/wordsmith

(Adopted from Reppen, 2010, pp.85-87)

Sketch engine (Figure. 1) for any word or a phrase, displays a concordance that lists example sentences drawn from a special text corpus, crawled from the World Wide Web, which has been cleaned of spam and includes only high-quality texts covering every day, standard, formal, and professional language.

Being a free corpus-based web tool that allows language learners and teachers find authentic sentences for specific target word(s), it consists of these major categories of word analysis: features, use, data, history, references, and external links. The number of displayed lines in a concordance is limited to 40. However, the frequency of the searched query in the reference corpus is indicated above the concordance as hits per million. This makes it a user-friendly tool search engine among the overstated providing so easiness and enough corpus examples to get acquainted with this new way of approaching foreign languages especially English now positioning itself as the core means of communication either physical or virtual across various contexts.

The present study aims to highlight the practicality of implementing among a range of tools as displayed in the SkELL Interface once you log in: Word Sketch, Thesaurus, Wordlist as innovative tools for a good start in exploring corpora, text, language; all of these products of social behavior(s) across different registers (see Figure. 2). It addresses the use of relia and authentic texts that guide autonomously language learners undertake, explore and reach to conclusive remarks to any relevant issue-specific empirical study of the English language.
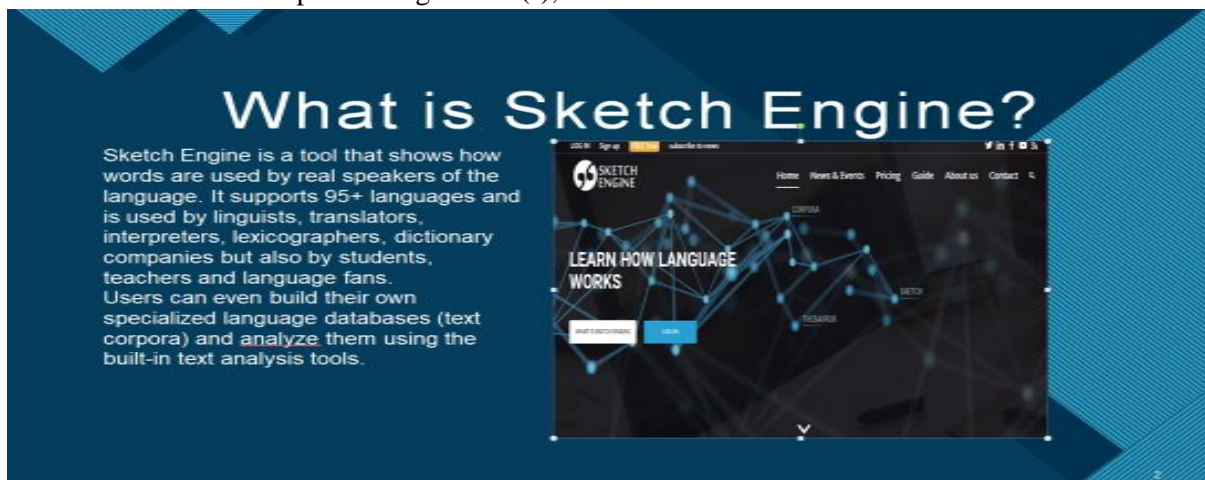
Figure. 1 Sketch Engine Interface
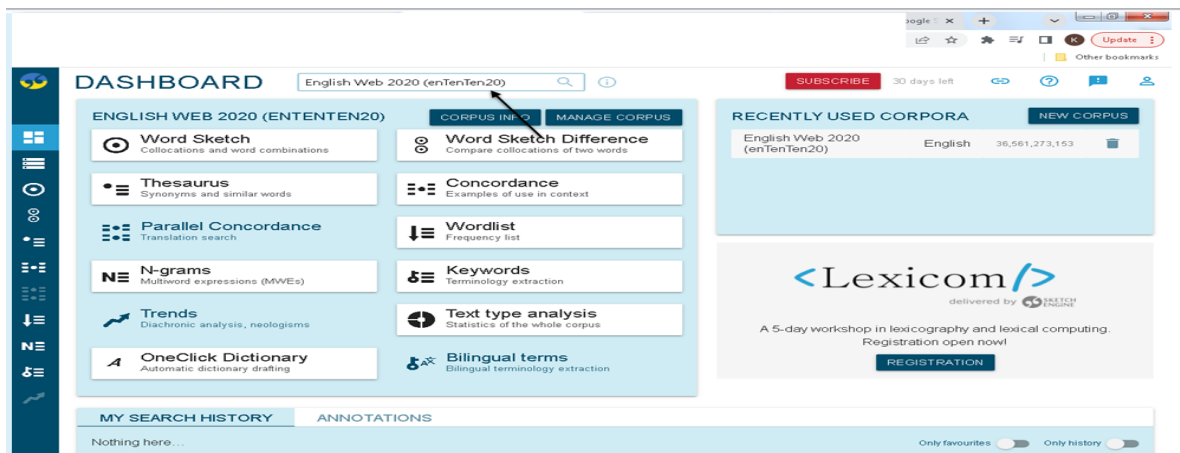
Figure. 2   Sketch Engine Tool Interface

3. SKETCH Engine: Generating language tasks through its Tools

In SkELL you can either look for at examples of word use or find out the word's collocates. It provides up to 40 examples of the use of the keyword in complete sentences as opposed to concordance lines, which is one of the main features that distinguishes this search engine among a range of others in which present Key Word in Context (KWIC) Search with all its occurrences in separate, disconnected lines in different contexts in any chosen corpus.

At first you need to register so to have access to use freely this search for 30 days. Then you should select the language you aim to use for linguistic and foreign language acquisition (provided our case we select on English

language). Then before choosing the language on the right side of the page there is a quick start tutorial on how to start to use the program. You can watch it in order to use the program properly and easily.

3.1 Word Sketch

Start in by logging in and switch to the new interface. This tool provides considerable amount of linguistic information in terms of the targeted word's variance in dependence to contextual collocations, frequency use across registers as displayed by viewing the keyword in a larger context in concordance lines. The following Figures 3 & 4 is a clear evidence of the actual tasks that can be generated by using this available tool from the Sketch Search Engine.

Figure.3 Word Sketch interface
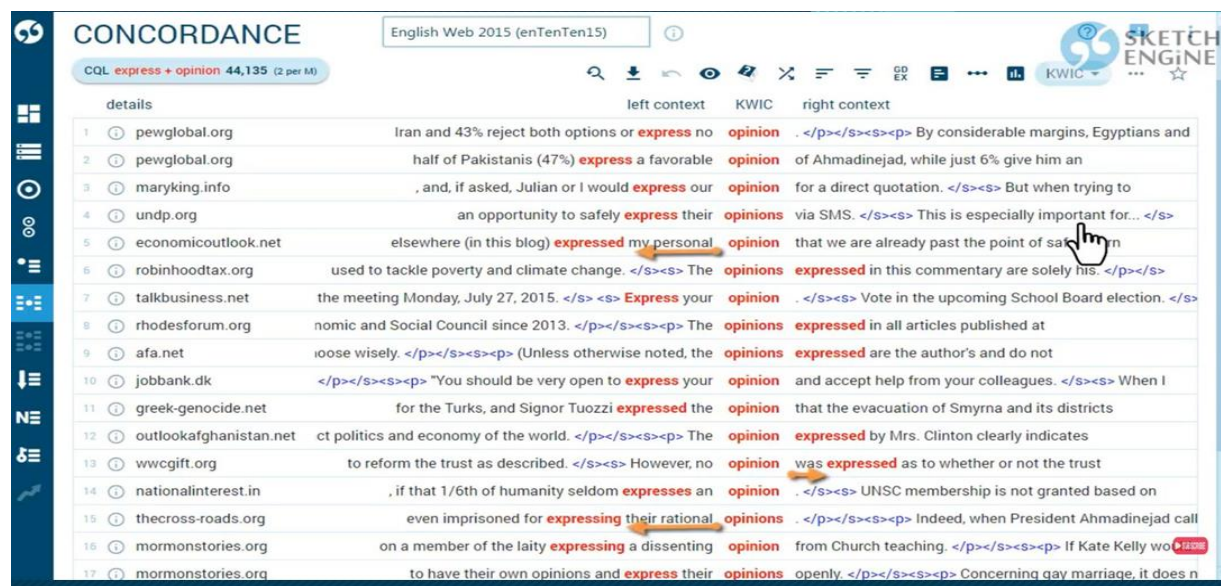
 Figure. 4 Word Sketch 'opinion' in a concordance.

Task 1. Type the word 'opinion' in the Word Sketch box and immediately this tool will identify as the data in Figure. 3 displays, the most important collocations of the targeted word.

Outcomes and analysis:

The program will display the data sorted by grammatical relations such as:

-   modifiers of 'opinion'. For example: majority/ expert/ humble opinion

-   nouns modified by 'opinion'. For example: an opinion poll, an opinion piece

-   verbs with opinion as 'object'. For example: express/ differ/ voice opinion

    Semantically having the connotation of differing opinions (information included on the column)

-   verbs with opinion as 'subject'. For example: opinions differ/ vary/ matter/

Task 2. You can check for the frequency of each of the displayed uses of the targeted word simply by clicking on the right side at the top of the above interface at the Change View Options icon.

You may sort words by frequency instead of typicality (Figure. 4)

Task 3. You may sort words by viewing the targeted word (KIWC) in concordance lines.

Task 4. Access visualizations. You can view data on the frequency of the word's collocates performing different roles as parts of speech and syntactically revealing different functions

Outcomes and analysis:

The corpus data that you may generate from the above tasks gives linguistic information on:

-   Frequency use

-   Collocation examples

-   Text type

-   Show scores

o The typicality score indicates how strong the collocation is.

o The higher the score, the stronger the collocation.

o A low score means that the words in the collocation also frequently combine with other words.

o Generating concordance this way might be easier than using the advanced concordance search.

- By viewing the targeted word in concordance lines you get to the conclusive remarks that Word Sketch is not just static of the words appearing in line before and after the keyword, but collocations can be identified correctly even if the latter appears at a certain distance.

3.2 Thesaurus

As can be noted from the following figures, Thesaurus is one of the tools that give you plenty of information right in an instant without having to carry any kind of dictionary or thesaurus to find the meaning of any encountering word in your language classrooms. It gives the possibility to see all the synonyms of a word. It suits to both types of student level either Basic or Advanced so as to experiment with the tool gradually and effectively. Thesaurus will usually work for these parts of speech only: nouns, adjectives, verbs, and adverbs.

Figure. 5 Thesaurus interface

Figure. 6 Change View Options of 'assimilation'

As can be seen from Figures 5 & 6 the tasks give all the information provided that they are generated completely automatically in real time and quite effortlessly. For example:

Task 1. Eliciting the synonyms of the academic word: assimilation (noun)

The provided outcomes pursue as the following:

   - Has the synonym words: segregation - 178, 178; globalization - 192, 963 ; colonization - 109, 208; unification – 85, 910; incorporation – 188, 771; modernization - 147, 513 ; ..etc.,

The score given for each synonym indicates the percentage of shared collocates.

Task 2. You can use the local menu to compare the synonym with the keyword:

- to display other synonyms

- to see examples of real text use

- to see collocations of the key word with other content words

Task 3. You can use the toolbar to modify the search criteria:

- To download the results

- Display visualization

- Display the search criteria

- Add the result to your favourites for easy access to next language class from the dashboard

Outcomes and analysis:

- The thesauri are useful because no manual work and no reference book or material is involved.

- The list of synonyms that is generated for any specific-targeted word in the language and corpus selected for linguistic investigation provides a sufficient number of occurrences found in the corpus.

- Moreover, synonym lists can be generated even for rare words which would not be included in traditional thesauri.

3.3 Wordlist

The Wordlist tool is used to generate frequency lists of all kinds: lists of words, lemmas, nouns,   verbs, tags, words containing or not containing certain characters

etc. The corpus we intent to use with the Wordlist tool is British National Corpus (BNC).
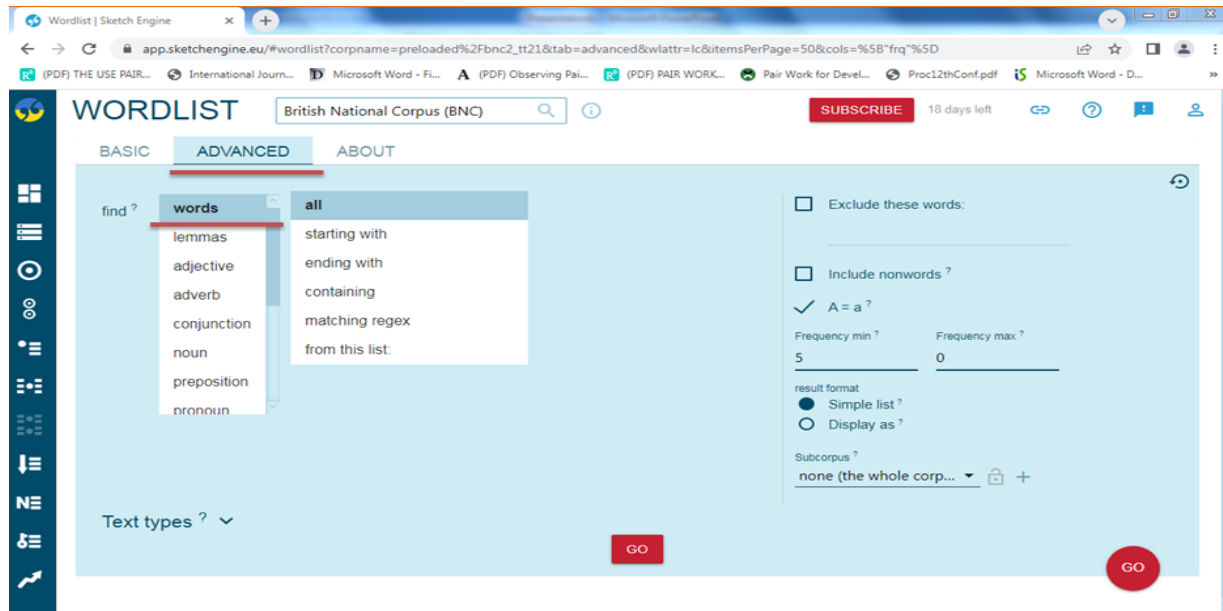

Figure. 7 Interface of Wordlist tool

Notably the above figure gives the proper interface that Wordlist display to both: Basic or Advanced options. We can click on a list word forms termed 'lemmas'. A lemma is a group of word forms that are related by being inflectional forms of the same base word. The lemma is usually labelled by that base or stem. So as McEnery & Hardie, (2012) for instance, explicate the significance in English of the words: destroy, destroys, destroying and destroyed that are all parts of the verb lemma destroy; each of which will be counted separately in annotated texts. But the noun destruction is a separate lemma, because it is related to destroy by derivational rather than inflectional processes.

Figure. 8 Interface of Wordlist Tool

Figure. 9 View options: Concordance 'discipline'

Task 1. More complex options are available on the advanced tab. You can write the word you want in the Search Box and observe input of its frequency.

The search result show word forms and it can be noted that the words "discuss" and "discussed" are treated as two different items. A frequency wordlist is simply a list of all types in a corpus with the number of occurrences of each type. The list as it is noted is displayed in frequency order, in alphabetical order, or in the order of the first occurrence in the corpus (Barnbrook, 1996).

Task 2. Use the local menu to go directly to the examples of the word in context or to display multiword expressions containing the word. The last two options are only available when lemma or part-of-speech is selected in the input form.

Task 3. Consider Figure. 8.

- If you click on the three dots right after the frequency values of each word, over the interface it will show up a new one with the options: Concordance, N-grams, Word Sketch, and Thesaurus. Select Concordance for this task.

- Examine the meaning and the behaviour of individual lexical items, and the pragmatic meaning of the corpus-presented phrases targeting the keyword "discipline" in (Figure. 9).

The straightforward questions that users may come up with are: What general assumptions about language do we infer in viewing concordance lines as a source of information?

The reply to this lies on the fact that concordance prioritise lexis and put emphasis on the association between pattern and meaning (both lexical and pragmatic).

Outcomes and analysis

   ✓ Word lists provide linguistic data on the frequency of a word or a phrase in different contexts giving so an important part of its description.

   ✓ Various word lists that are based to some degree on word frequency in a corpus exist especially in the English language teaching (ELT) context.

   ✓ Word lists are a good starting point for subsequent searches of individual items at concordance level and can be useful in the comparison of different corpora. But the limitation to this tool relies on the fact that the users need to be more proficient in apprehending conceptual notions typical of corpus language and consecutive forms of words in various interfaces of the tool itself (hereby referring to the Advanced search icon Tool on the web).

   ✓ Word lists can be generated to account for individual items or for recurrent sequences of two or more items (Hereby referring to the Basic search Tool on the web).

   ✓ Lemmatized frequency lists all group together words from the same lemma. For instance, the words 'say' , 'said', 'saying', 'says' are all part of the lemma SAY .

   ✓ Lemmatization can be done manually using an alphabetical frequency list, or in an automated way. Different forms of the same lemma tend to vary significantly in terms of their overall frequency, with one particular form tending to be more frequent than others. Such observations can be drawn by viewing lemma in concordance, a tool which stands on its own on the dashboard of the Search Engine, yet we saw it intertwined in supporting the lexical analysis in terms of frequency and collocation behaviour of the targeted word.

4. CONCLUSION

Corpora are widely acknowledged to be valuable sources in:

   ➢ describing language,
   ➢ compiling authentic language tasks,
   ➢ fostering through corpus-based approach and corpus-driven approach learner-learner interaction and autonomous language learning on the web.

For this reason, the present study has found it challenging and contemporary to introduce corpora and specifically SkELL a de facto implementation in English language Teaching and Learning. Provided that, language learners are always demanding on new approaches, instruments and environments either physical or

virtual (in reference to the current global insecurities we are still facing) and at the same time consisting in the intertwining of the theoretical literature into their practical involvement a 'product' that can be observed, studied and come to satisfying outcomes in the end.

Precisely, a corpus is 'real language', stored electronically. The relia provided in it is decontextualized and must be re-contextualized in a pedagogic setting by: applying the proper theoretical sociocultural approach/es, implementing the right field-specific tools and under the guidance of a language instructor to make it real to the language learner. Since the line of action: Corpus – Corpus Linguistics is meaningless with its adjunct that is Technology, demanding practitioners should be alert to all the advance and novelties that this strong asset has brought about and in a galloping speed is contemplating all the specializations of human endeavour.

Language teachers should guide learners towards what is novice in educational technology. Relying on the fact that they are keen on technological devices, the intertwining of new sociocultural approaches with technology integration as the case of Sketch Engine for Language Learning implemented with the researcher's ELT Master students will no doubt report successful. Yet, the implementation of SkELL and its tools: Word Sketch, Thesaurus, Wordlist used to analyse the linguistic input and output that corpora furnish across registers does not mean to solve all the language teaching issues. What is important is to take things one at a time, starting from basic knowledge to proceed to more conceptual and advanced mastery of Corpus Linguistics. If you know how to and where to use the tools, you can find the key to several complexities encountered on the web to frame foreign language teaching and learning just

as technology development wraps human existence.

5. References

1. Barlow, M. (1996). Corpora for theory and practice. International Journal of Corpus Linguistics, 1(1), pp.8-25.
2. Barnbrook, G. (1996). Language and Computers: A practical introduction to the computer analysis of language. Edinburg: Edinburg University Press.
3. Cobb, T. (2000). The Compleat Lexical Tutor [Website], developed by Tom Cobb of University of Montreal, Canada (Cobb, http://www.lextutor.ca/conc/eng/).
4. Friginal, E. (2018). Corpus linguistics for English teachers. New Tools, Online Resorces, and Classroom Activities. Routledge
5. Johns, T. (1986). Micro-Concord: A language learner's research tool. System, 14, pp.144-156.
6. Karpenko-Seccombe, T. (2017). Academic Writing with Corpora: A Resource Book for Data-Driven Learning: Routledge, Taylor & Francis Group, London and New York
7. Leech, G. (1997). Teaching and language corpora: A convergence. In Wichmann, A. , Fligelstone, S. , T. McEnery, T. & Knowles, G. (Eds.), Teaching and Language Corpora (pp. 1– 23). London: Longman
8. McDonough, J., and & Shaw, C. (1993). Materials and Methods in ELT: A teacher's guide. Oxford, England: Blackwell.
9. McEnery, T. & Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice, p.245. Cambridge University Press.
10. McEnery, T & Wilson, A. (2001) (second edition). Corpus Linguistics. Edinburg University Press.
11. O'Keeffe, A. (2020). Language teaching: Data-driven learning- a call for a broader

research gaze. Vol.54, Issue 2. Published online be Cambridge University Press: 21 July 2020.

12. O'Keeffe, A., McCarthy, M. j., & Carter, R. (2007). From corpus to classroom: Language use and language teaching. Cambridge, UK: Cambridge University Press.

13. O'Sullivan, I. (2007). Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy. ReCALL, 19(3), pp.210-245. CrossRef Google Scholar

14. Oztekin, E. & Candan, E (2019).The role of corpora in language teaching. (pp 327-343) In Research trends in English Language Teacher Education and English Language Teaching. Eksi, Y.G., Guerra, L., Werbinska, D.and Bayyurt, Y.(eds):University of Évora.

15. Pérez-Parades, P. (2019). A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011-2015. Computer Assisted Language Learning.

16. Available at: https://doi.org/10.1080/09588221.2019.16 67832 Crossref Google Scholar

17. Reppen, R. (2010). Using Corpora in the Language Classroom. In J. C. Richards, (Eds.) Cambridge: Cambridge University Press.

18. Römer, U. (2008). Corpora and language teaching.

19. Tonini-Bonelli, E (2001). Corpus Linguistic at Work. Studies in Corpus Linguistics, 6. xii. (pp. 65-80): Philadelphia, PA: John Benjamin Publishing Company

20. Trible, C. & Johns, G. (1997). Concordancing in the classroom. Harlow, UK: Longman. Google Scholar

21. Walsh, K. (2014). Engaging Uses of Instructional Technology. Posted in [EmergingEdTech]. Available at: www.emergingedtech.com/