International Journal of Arts Humanities and Social Sciences Studies
Volume 6 Issue 3 ǁ March 2021
ISSN: 2582-1601
www.ijahss.com

Using Corpus-Based Web Resources – How Can They Be Manipulated For Research in Language Patterns

Le Thi Giao Chi[1] , Nguyen Bich Dieu[2]

[1]*Faculty of Foreign Language Teacher Education, University of Foreign Language Studies - the University of Danang, 550000 Danang, Vietnam*
Tel: (+84) 236 3699 177 / (+84) 905 157 325
Email: ltgchi@ufl.udn.vn
[2]*Faculty of Foreign Language Teacher Education, University of Foreign Language Studies - the University of Danang, 550000 Danang, Vietnam*
Tel: (+84) 236 3699 177 / (+84) 934 863 568
Email: nbdieu@ufl.udn.vn

Abstract: *The introduction of Internet and the emergence of available web resources have facilitated endless efforts by language teachers to do research in language patterns. Web resources in general and corpus tools in particular have enabled large samples of language to be explored for better insights into the nature of language in use in all its forms and its uses. While corpora are normally assumed to be in the hands of lexicographers whose job is to inform dictionaries or grammar books, arguments may have arisen around why such end-users as language teachers and learners cannot make use of these innovative tools. This paper adds to this on-going debate by discussing approaches to using corpora as a reference point for language teaching and research. It shows potentials of manipulating common data-driven web tools for research in language patterns. It also explores some pathways for language teachers to examine aspects of language in use through authentic texts accessed via corpus tools. Results revealed from corpus search and generalizations made from manipulated data, and in this case noun and verb patterns and grammatical metaphor, in COCA and BNC can further showcase the inexhaustible implications of using web resources for enhanced language teaching and research.*

Key words*: web resource; corpus tools; corpus data; language patterns; grammatical metaphor*.

I. Introduction

The use of educational technology has opened up multiple pathways for innovations in language teaching and learning. Many a language teacher and education practitioners have counted on technologies and resources available online to constantly update their teaching content and enhance their instructional experience. In the arena of language research, web resources and tools have enabled large samples of language to be explored for better insights into the nature of language in use.

There has been quite a good body of literature done in the usefulness of Internet resources and tools for enhanced language teaching and learning (e.g. Kamba 2007; Son, 2011, Pim, 2013; Geladze 2015; Ahmadi, 2018). Kamba (2007), for example, examined the usefulness of the Internet as a tool for interactive learning, teaching and research. Son (2011) explored a number of online tools that can be used in second or foreign language teaching and learning. Pim (2013) recognized the prevalence of Information and Communications Technologies (ICTs) in supporting English language lessons, helping teachers to adopt innovative approaches in order to "meet the needs of the young technocrats growing up within an increasingly globalised world" (p.17). Geladze (2015) recognized the important role of modern informational technology and resources in innovating the teaching and learning process and in upgrading teaching on the scientific-methodological level. Erenchinova and Proudchenk (2017) acknowledged the use of Web resources in significantly improving the efficiency of learning a foreign language, enabling students to form and develop linguistic and communicative skills. Ahmadi (2018) further exhibited the impact of the Internet in advancing research, teaching and learning abilities as well as techniques that can be utilized by language teachers and researchers.

While it is important that the main goal of teacher as researchers is to search innovative ways for improving teaching and learning, knowledge on language patterns driven by large data size can really inform language teaching, thus adding to enhanced language learning. Without doubt, this depends on the new techniques and/or specific pedagogic technologies which can be realized by means of computer-assisted ways (Braun et al, 2006). This by all means brings to the fore the use of corpus technology, which is meant to inform language learning

and teaching with corpus-based observations revealing patterns of real language use, thus helping to uncover several problems of language in authentic operations. This paper, thus, adds to the on-going discussions of the multiple uses of internet resources, or rather, corpus-based web resources with case studies revealed for language teachers as researcher to explore aspects of language in use through authentic texts available online and/or accessed via corpus tools for research in language pedagogy.

II. Corpus Linguistics and Corpus-based Pedagogies – An Overview

Corpus linguistics is an increasingly popular field of linguistics, and according to Baker (2010, p. 93), it refers to "the analysis of very large collections of electronically stored texts, aided by computer software". In the position of McEnery and Hardie (2012), corpus linguistics deals with "some set of machine-readable texts which is deemed an appropriate basis to study a specific set of research questions" (2012, p. 1). Corpus linguistics is therefore a methodology, or an approach used to investigate linguistic phenomena rather than a sub field within linguistics (Ngula, 2018).

As its name implies, a corpus is a large 'body' of texts stored electronically. A general corpus is one that includes a variety of text types, ranging from written texts, spoken texts, or both, and very often it represents a national, regional or sub variety of a language. There are several general corpora of approximately a million words, such as the Lancaster-Oslo-Bergen (LOB) written corpus, and others of a much bigger size that include both written and spoken texts, such as the 100 million-word British National Corpus (BNC) and the over 450 million-word Contemporary Corpus of American English (COCA).

Figure 1. Introducing English Corpora

In fact, there has been a great consensus on the increasing availability of corpora and analysis tools that has made it easier to manipulate corpora in much wider ways, many of which are seen as relevant not only to language research but also to language pedagogy. With increasing interest in corpus technology for pedagogical uses, efforts have now been seen in integrating existing corpora, corpus methods and tools into teaching practice, and so have methods of exploring and expanding existing web tools geared for pedagogical research questions and goals. Thus, it can be said that corpus linguistics could revolutionize language teaching, by fundamentally changing the ways we approach all areas of pedagogy, including materials development, curriculum design, and teaching methodology.

III. Corpus-Based Web Tools For Language Research – Cases With COCA And BNC

All the functions of BNC and COCA can be useful in exploring language patterns. Why **SEARCH** is essential for getting to know the impressions of how language works, other functions such as KWIC, **COLLOCATES**, and **COMPARE** would be great to enable language teachers and students to study grammatical patterns, especially collocations, and to be able to explain why such patterns occur the way they are, and to get exposed to a great source of English grammar patterns as they are used in authentic contexts in different genres, spoken and written, academic and non-academic, fiction and non-fiction. Figure 2 shows a snapshot of some corpus functions.

Figure 2. Introducing Corpus Functions

Select a section: (sub-)genre or (set of) year(s). Click here for more examples of change over time.

*ize verbs in ACADEMIC	Past tense verb + *over* in SPOKEN
*ment in ACADEMIC	Nouns near *green* in 2000-2009
ADJ + *track* in NEWSPAPERS	Noun near *chair* in FIC
Hard + NOUN in MAGAZINES	Synonyms of *smart* in FICTION
Verbs in MAGAZINES-Sports	Nouns in NEWSPAPERS-Money
Adjectives in ACADEMIC-Medicine	Adverbs in FICTION-Movies

(Optional) Select a second (set of) section(s) against which to compare the sections chosen above

*ize verbs in ACAD vs MAG	Past tense verb + *over* in FIC vs ACAD
*ment in ACAD vs FIC	Nouns near *green* in 2005-2015 vs 1990-1999
ADJ + *track* in NEWS vs SPOK	Nouns near *chair* in ACAD vs FIC
Hard + NOUN in MAG vs ACAD	Synonyms of *smart* in FIC vs NEWS
ADJ in NEWS-Money vs NEWS	Nouns in MAG-Sports vs MAG
ADJ in ACAD-Medicine vs ACAD	Phrasal verbs with up 2005-2015 vs 1990-1994

These corpus-based web tools can also be used for preparation of paper-based grammar materials. In fact, language teachers and learners can use BNC and/or COCA, especially with the functions of Collocates, KWIC and Compare to find ample examples of real English language in use with confusing words like '*come'* vs. '*go'*, '*bring'* vs. '*take'* or '*small'* vs. '*little'*, and so on to design grammar exercises or quizzes for students. **COLLOCATES** can also provide insight into repeated grammar patterns that help teachers explore rules of grammar and that enable teachers to explain these rules to students with confidence and ease once real examples from real contexts are available.

BNC and COCA could also be great tools that facilitate self-regulated and collaborative learning. With a driving question to know about the grammar rules of some particular language phenomena like subjunctive or zero-conditionals or patterns with the pseudo subject 'it', etc., teachers can promote project-based learning even in teaching grammar and students are encouraged to do their projects, exploring COCA to find answers to their 'driving questions' or an assigned case of grammar, making notes, taking examples, comparing the use of grammar in different registers and prepare a presentation or a report using results from corpus-based tools presented in charts. There are huge applications for teaching grammar and vocabulary but what is important is that students need sufficient scaffolding from teachers to be able to explore the tool to their fullest extent.

IV. Corpus-based Web Resources and Research in Language Patterns – Some Case Studies

4.1. A Case Study with SPEND Patterns

In order to prepare for the material development, I did some research into the existing syllabus to find the topic which is considered as relevant to my students in the context of learning integrated language skills. As the communicative language teaching approach is preferable, I found that some practice of grammar in use would be relevant for students to lead in more meaningful communication about the topic. As Plans and Holidays is one of the topics in the course of General English, I noted it down and browsed the internet resources as suggested for this activity. I could see some relevance in the use of grammar to talk about holiday plans and the EVERYDAY GRAMMAR available at Learning English.voanews.com especially with the TWO -ING WORDS IN A ROW.

I then visited another resource suggested – American English Facebook (See Figure 3) and found an interesting match with a section on SPECIAL EXPRESSIONS + VERB + -ING with certain verbs like '*spend'*, '*waste'*, '*have a good/bad time'* to be used with this pattern. As these verbs can be used to describe how people spend their time and to plan a holiday, I decided to make some notes of the instances given on the webpage and design some meaningful practice for my students with gap-fill, sentence completion task before proceeding to some more communicative practice whereby students can make good use of the verb patterns to talk about holiday plans.

Figure 3. Screenshots of special expressions with Verb + ING

To further explore language patterns with verbs denoting holiday plans, I explored COCA and retrieved several authentic instances to adapt into the materials that I could design for the controlled practice of some grammar points before my students can proceed with more productive tasks – discussions about holiday plans. Students can then be given the links to access http://corpus.byu.edu/coca/ and explore the use of some key words with -ING used in real context. This can facilitate student autonomous learning and thus clear guidance, or instructions are necessary. For example, students can be asked to note down some special patterns from the concordance lines extracted from COCA when they search one of these key words: '*spend'*, '*waste'*, '*consider'*, '*try'*, etc. Figure 4 shows a snapshot of concordances with *spend*, for example, revealed in COCA, indicating the patterns in which the verb is used and the similar or different contexts it occurs. This enables students to examine the linguistic contexts whereby the word is used while being informed of the particularly distinctive features characteristic of the word.

Figure 4. Concordances of SPEND in COCA

4.2. A Case Study with TRY patterns

In order to identify the patterns of language of ***try*** as its part of speech varies, the *List* and *KWIC* functions in COCA are exploited to compare the frequency of the word ***try*** and observe their concordance lines. With the *List* function, I put the word try in the SEARCH box. By clicking FIND MATCHING STRINGS, the frequency comes up to 138 333. By clicking the word ***try***, it reveals the authentic examples with ***try*** in the concordance lines and the real contexts in which the word occurs. When tagging POS to the word with verb.all – try.[v*], and with noun.all – [try.[nn*], **133 736** and **4589** occurrences come up on the screen respectively, which shows that ***try*** is more often used as a verb than as a noun.

With this initial finding, I continued to explore the word ***try*** used in real contexts, and this throws some very interesting insights into the common patterns or phrases whereby the word is used. When used as a verb, ***try*** can function both as an intransitive verb or a transitive verb. Being an intransitive verb, ***try*** is found predominantly with the adverb **again** as in '***Try*** *again'*, '*Please try again' 'Why don't you* ***try*** *again'* or **out** as in '*OK. Come* ***try*** *out'* or *'****try*** *out again' '****try*** *out over the weekend/underwater/for the next season'.* When acting as a transitive verb, ***try*** needs an object be it *a noun, a noun phrase* (e.g. *try innovative ideas*), or a gerund. Interestingly, I found out that the phrase ***try out*** has been used quite often. As I searched the entire phrase, I could find **1207** occurrences of ***try out*** in the corpus with plenty of interesting examples with concordance lines - ***try out*** **+ N/NP** (e.g. *try out the position / new powers / different ideas / the new fundraising platform*; or *try out someone that knows how to really find the love in any situation*, etc.) or ***try out*** **+ V-ING** like the following

'If only the youthful years were set aside for the children to **try out** being a boy one day, a girl the next, an angel the day after, a monk or an astronaut on the following days' (Context from COCA)

When used as a noun, ***try*** is often preceded by a determiner (e.g. *a* ***try***) occurs in some common patterns such as

- give it a **try**
- worth a **try**
- **m**ake a **try**
- **g**ive one more **try**
- attributive adj + **try** (a real **try** / nice **try** / good **try** / his best **try** / his latest **try**)
- on his first **try** / on the second **try**

The KWIC function with the two most common patterns with **try** where the frequencies can reveal which of the two is more prevalent in the corpus. While ***worth a try*** occurs up to 256 times, its counterpart ***give it a try*** is even much more commonly used in the corpus, especially in spoken discourses (619 occurrences). Also, the results from KWIC function of COCA show that ***worth a try*** often occurs with the verb 'be' with or without adverbs showing certainty like *certainly* or *definitely.*

It is	worth a ***try*.**
It was	
It may be	
It might be	
It's certainly/definitely	

While patterns ***It + BE + worth*** a try are used predominantly in fiction, patterns with ***give it a try*** are then found in more various registers: fiction, news, spoken, and magazine. When being used in spoken domain, ***give it a try*** often comes with *Please / I'll / Why don't you* [give it a try] to show ways of giving advice or making suggestions.

Figure 5. Concordances of 'worth a try' in COCA

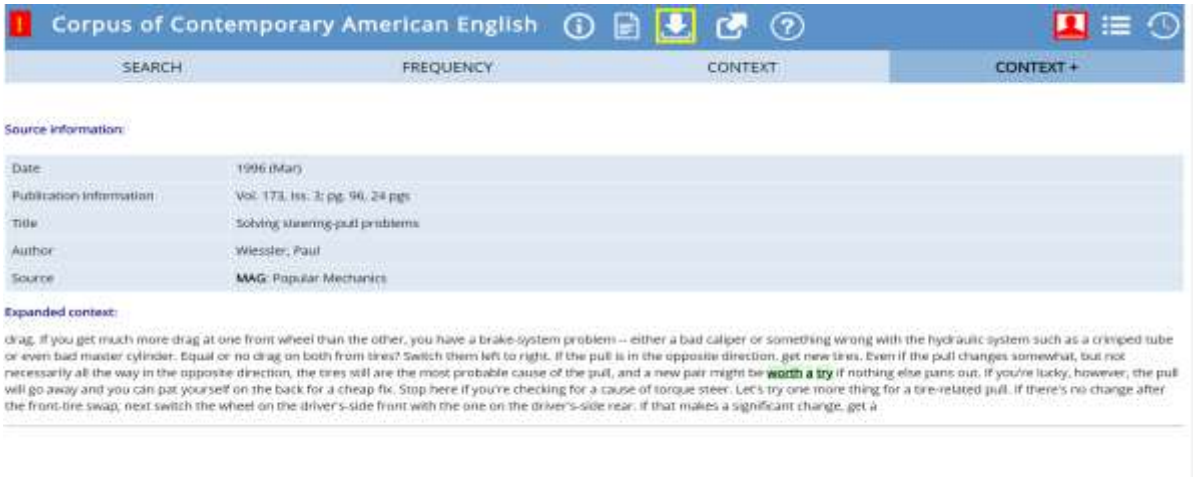


Figure 6. A sample context for 'worth a try'

4.3. A Case Study with Confusing Words – LITTLE vs SMALL

To examine the differences in language use between synonyms or near-synonyms, the COMPARE function in COCA/BNC is used. For example, results from COCA show different representations of these two (near-)synonyms with **little** being 1.84 and **small** being 0.54. Although both words mean '*small in size, amount, or degree'* their occurrences in real contexts of situation can reveal some differences. When **little** is used as an adjective in attributive position in a sentence or a phrase like *a **little** sister/ a **little** bastard*, it expresses some *affectionate* attitude towards the subject being mentioned. **Small**, on the other hand, merely shows the meaning of 'small in size'. *My **small** brother* means 'he is small, smaller in built than average'. **Little** has more diverse syntactic roles, functioning as a **determiner** preceded by a pre-determiner 'a' like ***a little** baby/ **a little** girl*, or as an **adverb** modifying an adjective like *feel a **little** annoyed / hard.*

With the display of COCA, collocates with **little** and **small** are shown in two lists: W1-LITTLE vs. W2-SMALL and V1-SMALL vs V2-LITTLE. This enables the noting down of most common collocates with either W1 or W2 or with both. The word BIT, for example, is found in collocates with both **little** and **small. Little** in **a little bit** is found with a frequency of 192.9 or 40 889 occurrences (*feel a **little** bit of annoyance/ a **little** bit arrogant/tired/crazy/…*). *Bit* is also found with **small,** though with lower frequency – 105.1 or 212 occurrences. The word BUSINESSES is also recognized to go with both **little** and **small**, though of course, with the meaning of 'limited in the amount of activity' it collocates with small up to 3365 times. The phrases ***small** businesses* are common in today business world with 3741 occurrences (See Figure 7). Similar phrases of this meaning are ***small** producers / employers/farmers/donors, etc. Small* can also be found in collocates with ***small** sums, **small** scopes, **small** tumors* whereby its central meaning is 'small in size'.

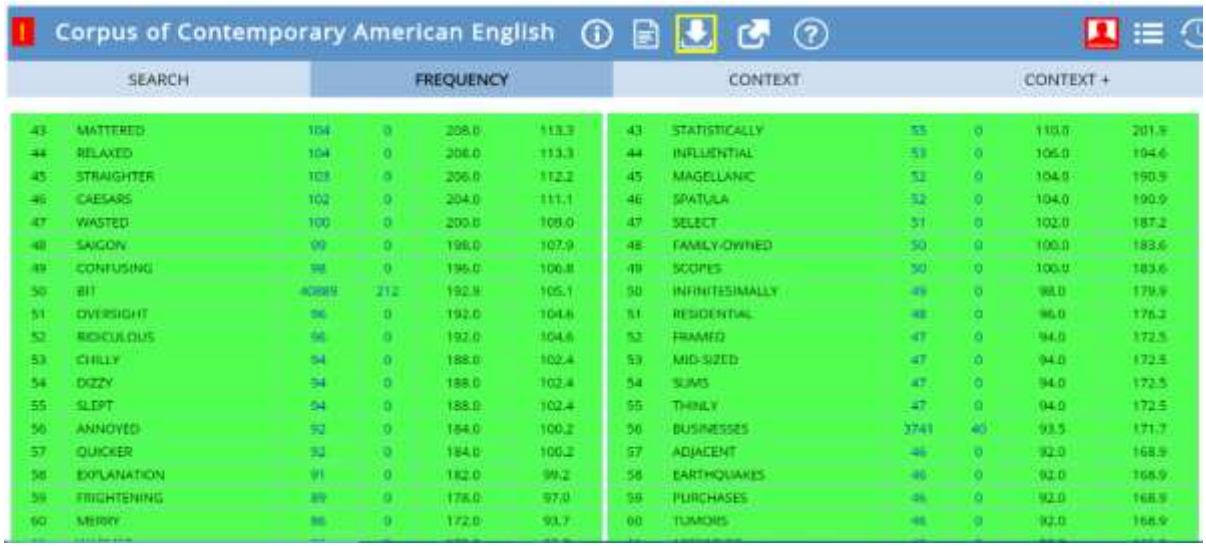


Figure 7. Collocates with 'LITTLE' vs 'SMALL' in COCA

This COMPARE function thus helps us explore words that collocates with one rather than the other word. With the distinctive function as an adverb that precedes an adjective or a comparative (*a* ***little*** *guilty/younger/louder/faster/slower, …*), **small** doesn't reveal any matches. However, in cases where the meaning conveyed is 'limited in size/amount/scope or smaller than the average size' we can see more collocates with **small** like small budgets (27)/amounts (1126)/quantities (283)/percentages (1231)/proportions (290), …. Interestingly, SAIGON, the name of the biggest city in Vietnam occurs 97 times in **Little Saigon**, which may indicate reference of a small Vietnamese community in the US. Similarly, the proper name **Little League** occurs up to 1445 times in the News and Spoken texts of COCA with *Little League Game* or *Little League Coach.*

Figure 8. Collocates with 'LITTLE' vs 'SMALL' in BNC

Some interpretations into the use of **little** versus **small** in its collocation can be gained here. While the left section shows chances that certain words can go together with ***little*** versus ***small*** with corresponding occurrences and percentages, the right section shows the opposite. It is thus interesting to find that we can see **small** in ***small*** *politics*, ***small*** *organisations*, ***small*** *treatment*, ***small*** *test*, ***small*** *techniques*, ***small*** *procedures*, etc. whereby the number zero in the column denotes that there is no such case of collocates with **little**.

4.4. A Case Study with Nominalisations

With the SEARCH function, both COCA and BNC can reveal results regarding frequencies, context, and also collocations of nominalisations as grammatical metaphor in authentic contexts. With the trial in the search box for such nominialisations ending with -MENT, results extracted from the COCA tool can throw up

most common collocations with the search word being identified with supporting frequencies ranked in the descending order. While such nouns as 'department', 'moment', 'element', 'segment' are clearly not the case, nominalisations like 'develop*ment*', 'treat*ment*', 'move*ment*', 'assess*ment'*, 'manage*ment'*, are at the top of the list with frequencies being shown in Figures 9-10.

Figures 9-10. Frequencies of nominalisations with -MENT in COCA

The CONTEXT function in COCA/BNC allows also a deep look into the real contexts where each of the nominalisations occur and also the various contexts of the same nominalization, in this case 'growth' to see how and what collocates with it. Interestingly, results from the tool inform certain patterns whereby nominalisations deriving from verbs are seen to collocate with other nouns or prepositional phrases premodifying or post-qualifying the nominalized noun in the whole string with nominalisations. The complex semantic wrappings indicative of these nominalized segments like *wireless data* ***growth****, the rate of wage* ***growth****, the* ***growth*** *dreams of some developers, in better* ***growth*** *numbers, the unprecedented viral* ***growth****, a huge area of* ***growth*****.** Figure 11 shows the context of *growth* and its concordance lines in the register of magazine in COCA.

Figure 11. Concordance lines of nominalisation 'growth' in COCA

The nominalised constructions deriving from an adjective or de-adjectival nominalisations can also be extracted, such suffixes as *-ity*, *-ness* for example, denoting *the status or the feature of being the adjective.* Concordances from COCA or BNC can throw further insight into the operationalisations of these nominalised segments which are found to collocate with other nouns or nominal groups or prepositional phrase with *of* post-qualifying the nominal and predeterminer *the*:

While Shields believes in the authenticity and continuity of the craftsman and does not interrogate craft masculinity as a social construct

Figure 12 illustrates the concordance lines of de-adjectival nominalised segments with ***-ity*** in COCA while Figure 13 similarly prescribes the window to the context of concordance line 16 when a click is made on its context. With such knowledge, language teachers as researcher can probe to the depth of how language operates in its authentic context, thus informing students of the real language in use.

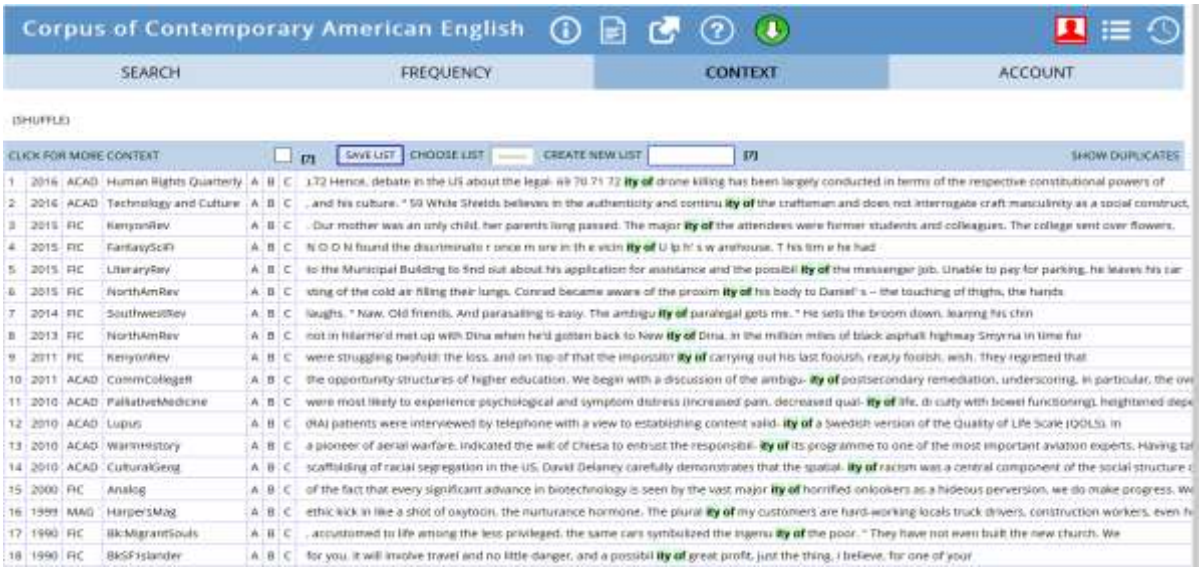


Figure 12. Concordance lines of nominalisations with -ITY in COCA

*Figure 13. Context of **plurality** as a de-adjectival nominalisation in COCA*

Nominalised constructions with **-ness** are searched in similar ways and results reveal showing the frequencies, contexts of situation and the wide variety of genres which are characteristic of such language patterns. Accordingly, the patterns with de-adjectival nominals with -NESS are strongly indicative of the *-of*-phrase being in the post-qualifying function. As Figure 14 illustrates, the nouns following *of* in the nominalised constructions can be coded as the agent that possess the features or quality of the adjective from which the nominalization is formed. For example, *the sheer richness of the painting* (concordance 22) must come from its congruent segment of 'the painting is completely rich' or *the slowness of her progress* (Concordance 41) being 'her progress is slow'.

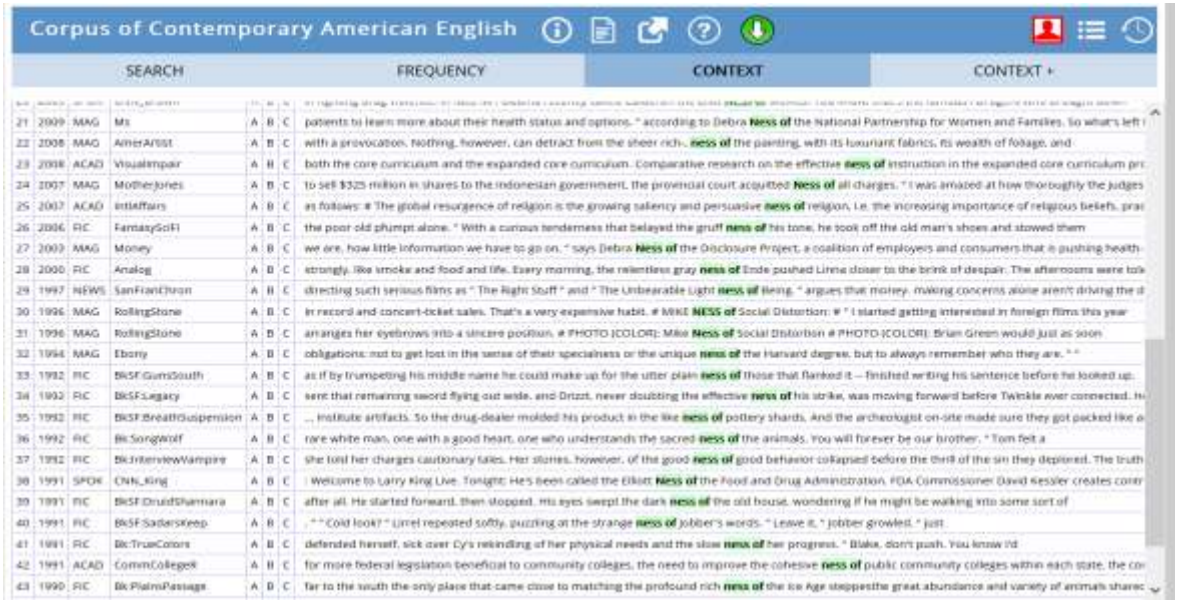


*Figure 14. Context of **plurality** as a de-adjectival nominalisation in COCA*

Further concordance lines from COCA can also be taken from various years of media publications with real contexts being given for nominalisations with -ness (See Figure 15).

he didn't see what harm it could do to treat Clair with a little extra fond**ness** and leniency - *Virginia Quarterly Review*

And when I walked through the tinted-glass doors of the cool, cave-like computer lab into the warmth and bright**ness** of the school's courtyard, how satisfying to hear that familiar voice from the oaks or the wisteria trained on a trellis overhead. **Southern Review, Vol 15, Issue 1 2015*

It was less sad than Jill had feared, the empti**ness** of the place more peaceful and powerful. *Ploughshares, 2015, 41 (2)*

With that Clair found him there, standing at the edge of the tank. Her face initially reg istered surprise but soon deepened into kind**ness**. " I didn't think you were going to make it - *Virginia Quarterly Review 2015, 91-3.*

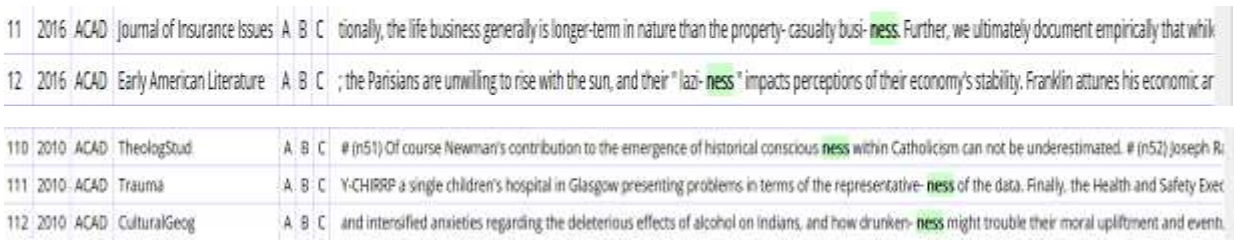


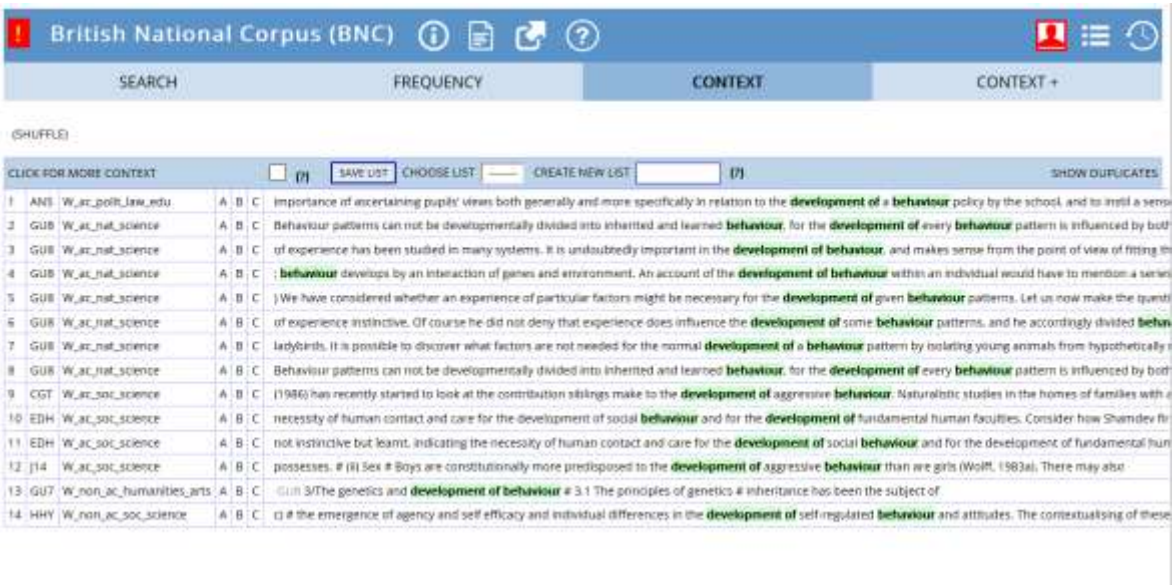
Figure 15. Context of de-adjectival nominalisations with -ness in COCA

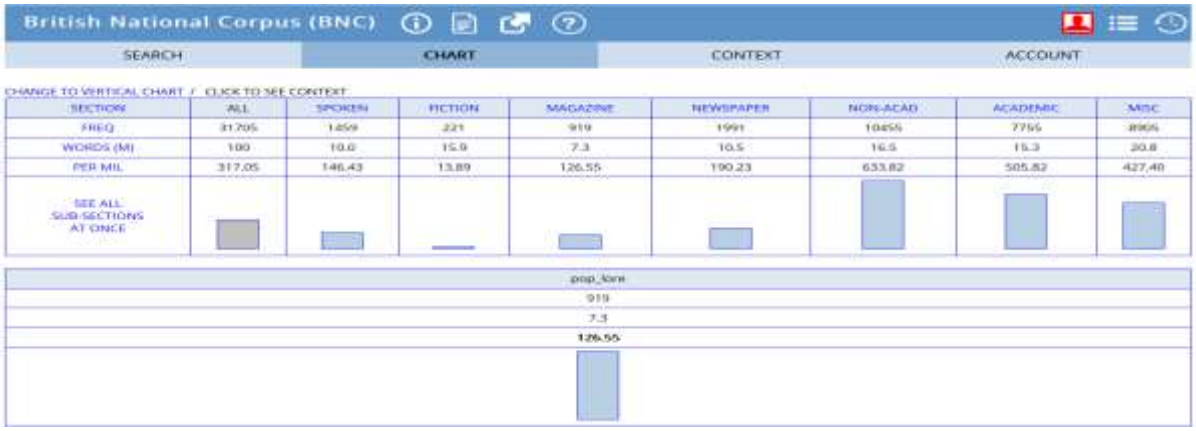
The COMPARE function is also used to compare nominalisations as near-synonyms. Two synonymous nominals like *development* and *growth* are put in the COMPARE box, and cases showing collocates with either

or both of these words are shown. For instance, collocates of *growth* with *love* is found rare, and still rarer with *development*. In fact, no such case throws up in the search.

*Figure 16. Some concordances with '**development**' and with '**growth**' in BNC*

Corpora like COCA and BNC also allow comparison made into the patterns of language used in different registers. Occurrences and percentages corresponding to each of the genres where the search word or phrase are examined can be shown in nicely-presented tables, providing interesting information about the features of any language patterns (Figures 17-18).





*Figures 17-18. Concordances, occurrences and percentages of '**development'** in BNC*

V. Conclusion

With the main goal of teacher as researchers being established which is to search innovative ways to improve language teaching and learning, knowledge on language patterns driven by large data size is thus important to really inform language teaching, thereby enhancing language learning. Without any shadow of a doubt, all teaching transformations require that new techniques and/or specific pedagogic technologies be incorporated in computer-assisted ways and by means of web-based resources and tools with we are to keep up with the pace of such development.

With more and more teachers as researchers probing into the depth of language in use with the aid of corpus-based web tools, the future is not far away when increasing authenticity and practicality is imported into the classroom. Hopefully, minor case studies as presented in this paper are of some relevance to the on-going endeavor for bridging the gap between corpus linguistics as a sub-field of linguistics and corpus-based studies driven by web resources as toolkits for further explorations into language patterns for enhanced language pedagogy.

REFERENCES

[1]. Ahmadi, M. R. (2018). The Use of Technology in English Language Learning: A Literature Review. International Journal of Research in English Education, 3 (2), 115-125.
[2]. Baker, P. (2010). Corpus methods in linguistics. In L. Litosseliti (ed.), Research Methods in Linguistics (pp. 93–113). London: Continuum.
[3]. Braun, S., Korn, K. and Mukherjee, J. (2006) Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods. English Corpus Linguistics, Vol 3.
[4]. Diakou, M (2015) Using information and communication technologies to motivate young learners to practice English as a foreign language in Cyprus. Jaltcalljournal, Vol. 11, No.3, pp. 283–292.
[5]. Erenchinova, E. and Proudchenk, E. (2017) Advantages of WEB resources use in learning process. Advances in Social Science, Education and Humanities Research (ASSEHR), Volume 97, 103-108.
[6]. Geladze, D. (2015) Using the Internet and Computer Technologies in Learning/Teaching Process. Journal of Education and Practice, Vol, 6, No.2, 67-70.
[7]. Kamba, M. A. (2007) The Internet as a Tool for Interactive Learning, Teaching and Research: Nigerian Experience. IMCL Conference.
[8]. McEnery, T., & Hardie, A. (2012). Corpus Linguistics: Methods, Theory and Practice. Cambridge: Cambridge University Press.
[9]. Ngula, R. S. (2018). Corpus Methods in Language Studies: Perspective on Conducting and Reporting Research in the Humanities, 205-223.
[10]. Pim, C. (2013). Emerging technologies, emerging minds: Digital innovations within the primary sector. In G. Motteram (Ed.), Innovations in Learning Technologies for English Language Teaching (pp. 15–42). London: British Council.
[11]. Son, J.-B. (2011). Using Web-Based Language Learning Activities in the ESL Classroom. International Journal of Pedagogies and Learning, 4(4), 34-43.
[12]. The British National Corpus (BNC)
[13]. Contemporary Corpus of American English (COCA)
[14]. The American Facebook Page.