Received: September 15, 2017 | Reviewed: October 1, 2017 | Accepted: November 1, 2017

UDC 811.111(072):159.9 | DOI 10.5281/zenodo.5153452 | Research Paper Citations

PRACTICAL USES OF CORPUS TOOLS IN THE PSYCHOLOGY CLASSROOM

Liliana COŞULEAN
Assistant Professor
(Alecu Russo State University of Bălți, Republic of Moldova)
liliana.golubenco@mail.ru

Abstract

In the present article I would like to introduce both teachers and avid learners of English into the wonderful world of Corpus Linguistics as a useful yet relatively novel tool of linguistic research and a resourceful method in ESAP teaching. Technology is introduced into the classroom by means of corpus tools. The setting is academic. The partakers are a group of psychology students with A2 to B1 knowledge of English in their first course of studies. Above all, the purpose of such undertaking was to introduce Data-Driven Learning into the academic classroom in order to enhance learner autonomy, incite curiosity in learners and to switch from prescriptive views of language to descriptive views, all these becoming possible due to well-balanced learner corpora of millions of authentic English texts across various registers of language.

Keywords: *Corpus software, data-driven learning, learner autonomy, concordances, query, lemma, cross-register analysis, frequency list, natural language processing*

Rezumat

Articolul dat este o încercare de a prezenta atât profesorilor, cât și celor care învață limba engleză, lumea minunată a Corpusului Lingvistic, ca instrument util, deşi relativ nou, de cercetare lingvistică, dar şi o metodă plină de resurse în predarea limbii engleze în scopuri academice. Tehnologia este introdusă, la oră, prin intermediul instrumentelor de corpus. Cadrul este academic. Participanții sunt un grup de studenți la psihologie cu cunoștințe de limba engleză, nivel A2-B1, în primul an de studii. Mai presus de toate, scopul unei astfel de lucrări a fost de a introduce învățarea bazată pe date în sala de clasă pentru a spori autonomia elevilor, a incita curiozitatea, dar și pentru a realiza o trecere de la abordarea prescriptivă a limbajului la cea descriptivă, toate acestea devenind posibile datorită corpusurilor balansate de milioane de texte autentice în limba engleză, raportate la diverse registre ale limbii.

Cuvinte-cheie: *Corpus softuri, învățare bazată pe date, autonomia elevului, concordanțe, lemă, analiza registrelor, listă de frecvențe, procesarea limbajelor naturale*

Ever since its beginning, dating back from 1960s when it was conducted by a small group of modest enthusiasts, Corpus Linguistics has been facing a great deal of criticism from influential linguists and language acquisition connoisseurs. Apparently, at a certain point, Chomsky's criticism (Aarts, 2000, pp. 5-15) was one of the most important impediments in the growth of corpus studies and at that point had an immediate impact: the rationale for deep research was regarded as insufficient and even nonsensical. The point was that unlike many of his predecessors, Noam Chomsky did not share empirical views asserting that learning a language is an imitation of some sort. Conversely, he argued that the knowledge of a human language is something innate, that there must be something in our mind to make the language acquisition possible. “One of the big insights of the scientific revolution, of modern science, at least since the seventeenth century… is that arrangement of data isn’t going to get you anywhere. You have to ask probing questions of nature. That’s what is called experimentation, and then you may get some answers that mean something. Otherwise you just get junk” (Noam Chomsky, apud Aarts, 2000).

At that point, Chomsky’s position urged the early researchers reevaluate their work. In the beginning of corpus research critics claimed quantitative limitations as it was considered impossible to process such a great amount of written text of several million words in length. About twenty years ago it was considered slightly possible, but nonetheless time-consuming. Maybe surprisingly, but today it is becoming increasingly popular. Presently, the utility of corpus studies is not questioned any more, most corpus enthusiasts disagree solely on whether it is a powerful approach, a methodology, a method or a technique.

In the last decades Corpus Linguistics has been very prolific regarding its most practical application in foreign language acquisition, more exactly in languages for specific purposes in academic environments. It is also true it has been infiltrated in all language-related disciplines providing most unexpectedly successful outcomes.

In the foreign language classroom, we should refer to an innovative approach to learning called Data-Driven Learning (DDL) which is actually gaining ground in foreign language teaching. While the core of most approaches lies in the teacher-guided learning and textbooks, in data-driven learning students are encouraged to treat language as a large data hub where they can carry out a great variety of discovery tasks on their own. In this context, many language teachers take high the principles of the DIKW pyramid- an underpinning concept which refers loosely to a class of models for representing structural and functional relations between the four levels: data, information, knowledge and wisdom. "Typically information is defined in terms of data, knowledge in terms of information, and wisdom in terms of knowledge" (EJ, pp. 163-180):

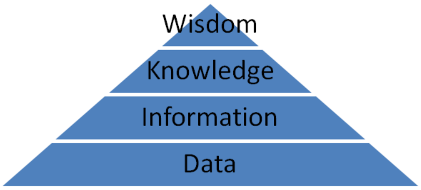

Table 1: *The DIKW Pyramid*

The DDL is characteristic of a pattern-based approach to grammar and vocabulary and a lexico-grammatical approach to language in general. Applied to corpus research tools, students explore authentic texts; they see how language changes and how it is used, they make generalizations and observe common patterns. Additionally, they learn how to obtain data from the source and how to interpret the original data. Ultimately, students make inferences and come to conclusions, they are able to use high order thinking abilities and they start to create knowledge. In data-driven teaching styles teachers would suggest a corpus of written texts, stored and sampled electronically, to be researched online or paper-based. It should be mentioned that corpus tools are especially designed for linguistic analysis; however the newest versions were created namely for classroom use. The primary learning objective is to bring authenticity to the classroom through exposing learners to authentic language.

According to Gabrielatos, work with a representative corpus is essential to the ability of recognizing patterns, in countries where the target language is scarcely spoken. It is perhaps the only way to have direct exposure to authentic language. The notion of "condensed language exposure" has emerged to denote a mixture of extensive and intensive reading strategies which are times more efficient than traditional reading. Extensive reading is regarded as an effective way of language learning because of exposure to real language use in natural contexts and in larger amounts than short texts and dialogues that are used when presenting new grammar or lexical material (Gabrielatos, 2005, p. 21).

Extensive reading is also regarded as an effective way to help language learners develop intuitions about language in use. After Gabrielatos (*idem*, p. 33), the importance of condensed learning does not imply that previous forms, extensive or intensive, should be dropped out, on the contrary, he suggests that the new approach should combine both forms. This way exclusively, learners can formulate and check hypotheses about language structure and use, as well as focus on particular language features. However, this task must be accomplished taking high the representativeness and the size of the sample; in other words, the collection of texts needs to fully represent the language use of the population under investigation and not be too large and complex for a small sample to reveal adequately to avoid over-generalizations based on inadequate or selective evidence.

Corpus Techniques Applied

From now, I would like to share some considerations on how corpus tools can be used in the English classroom at Alecu Russo Balti State University. The setting is academic; the trainees are first year students intermediate level of English, pursuing their Bachelor's degree in Psychology. Professional English is taught with specialization in Psychology.

The selection of corpus was made keeping high two main factors acknowledged by Douglas Biber (Biber, 1993, pp. 243-257) as important considerations in a work of this sort:

1) *size of the corpus* (including the length and number of text samples);

2) *range of text categories* (or registers) that samples are selected from.

Worth noticing that according to Biber (*idem*, p. 249), they are associated with "random errors" and "sample errors" and can threaten validity, i.e. the extent to which we can make generalizations from a sample for the target audience. He further claims that "random error occurs when a sample is not large enough to accurately estimate the true population; bias error occurs when the selection of a sample is systematically different from the target population it is intended to represent. Both kinds of error must be minimized to achieve a representative corpus" (*idem*, p. 247).

Thus, a teacher has a great choice of various corpora: large , medium and small, depending on several factors. The Brown Corpus and LOB Corpus, which are small by present-day standards and are explicitly structured to 'represent a wide range of styles and varieties' (Francis and Kuchera) and COCA which is considerably larger and can also boast a wide variety of registers. Projects such as the COBUILD Corpus, Longman/Lancaster Corpus, and British National Corpus (BNC) combine both emphases to a certain degree and can also be used. However, after several tryouts with students, I was inclined to take two of them: the COCA corpus[17], due to its simplicity of use, interactivity and wide-ranging features and the COBUILD corpus due to the explicit structure and representativeness for the British variant of English.

The trainees, who were introduced to the fundamentals of corpus-based analysis, from a scratch, were able to perform the simplest operations in language analysis with permanent hands-on the corpus: lexical, structural, lexico-grammatical, morphological patterns, collocations, frequency lists, etc.

Below, there are some basic corpus techniques applied and successful with students. Apparently, they may suffer some adjustments, can be either completed or attuned to group level, educational needs, settings and time limitations and used in a manner the teacher will choose individually. The examples engage extended search for the word "Psychology" in various activities:

- *Cross-register work.* The goal of this activity is to reveal that language behaves differently according to various registers. With proper analytical tools the trainees will find out not only the patterns themselves, but the extent to which some profession-based terminology undergoes changes in incidence across registers. As a result, students could single out the most

[17] https://www.english-corpora.org/coca/?b=x2&c=coca&q=22285649.

common registers where the word "psychology" appears. Thus, the ***academic*** register was far ahead the other registers such as Blog, Web, Magazine, News, Spoken, TV/ Music or Fiction which came last in the list:

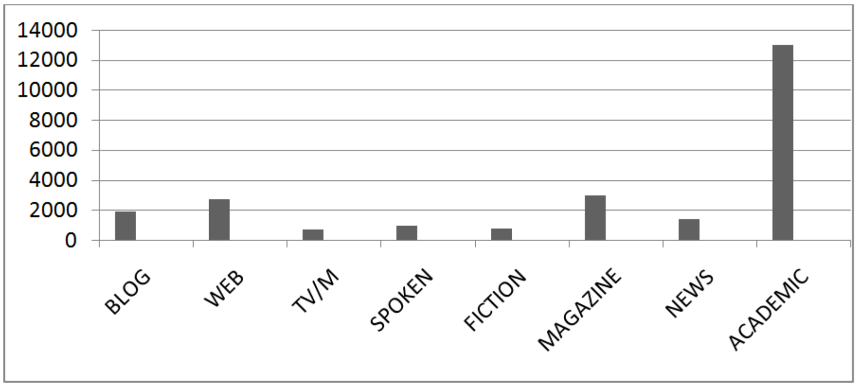

Table 2: *Word Frequency of the Word "Psychology" across Registers in the COCA Corpus*

In the next phase, students can discuss their results and compare numbers that show the discrepancy in usage frequency of the given word across registers. In the same manner, they would compare two or three different corpora so as to identify the register where the word *psychology* occurs most frequently and thus they can draw relevant conclusions, which is a sign of gaining higher level skills.

- *Diachronic search.* Tracking out when a certain word was most used in the history is a popular technique which brings enjoyment and raises curiosity in most students as they can observe the rise and the fall of a certain word throughout time. In larger corpora of this kind, a term comes in many different registers displaying different frequency results. Let's take, for example, the COCA corpus, where the results look as follows:

	1990-94	**1995-99**	**2000-04**	**2005-09**	**2010-14**	**2015-19**
FREQ	15780	19559	14658	13848	11589	10707
WORDS (M)	139.1	147.8	146.6	144.9	145.3	144.7
PER MIL	113.48	132.36	100.01	95.54	79.78	73.97

Table 3: *Diachronic Records for "Psychology" in the COCA Corpus*

- *Collocation work.* Collocations are just two words or more that fit together in a corpus of work more than chance could allow. All sort of corpus-based activities in the classroom with stress on collocations are very prolific. Students benefit from this type of corpus research as it gives them an insight into word combinability. The goal is to find out functional phrases, the utility collocations as they occur in natural languages.

The most common words that come before or after the noun psychology are certain verbs, nouns, adjectives or adverbs which in terms of frequency were singled out during this task:

Freq. no	+NOUN	Freq. no	+ADJ	Freq. no	+VERB	Freq. no	+ADV
2560	School	**1355**	Social	**247**	study	**96**	e.g.
2211	Journal	**1054**	Educational	**102**	exercise	**20**	i.e.
1758	Professor	**744**	Clinical	**91**	enroll	**20**	quarterly
1700	University	**416**	Human	**48**	graduate	**9**	Extensively
878	Sport	**408**	Cognitive	**41**	major	**3**	Hitherto
799	Student	**387**	Experimental	**32**	recruit	**2**	Experimentally
677	Department	**360**	Developmental	**23**	specialize	**2**	inwardly
540	Review	**358**	Applied	**12**	delve	**2**	Amok
514	Course	**356**	Evolutionary	**6**	bowl	**2**	Oft
504	Research	**249**	Introductory	**5**	minored	**2**	Clumsily
466	Science	**200**	Positive	**4**	school	**2**	biweekly
442	Degree	**189**	Professional	**4**	intern	**2**	authoritatively
434	Education	**154**	Associate	**4**	verse		
423	Study	**145**	Abnormal	**3**	underpin		

Table 4: *Hits for "Psychology" Collocations in COCA*

On completing this task, students will be able to make generalizations eliciting that nouns have the highest collocation ability, adjectives approximately twice lower, whilst verbs have collocation capacity ten times lower compared to nouns. Also, they will notice that last in the list come adverbs with very low collocation ability. During this activity it is suitable to let students make up their own sentences with collocates. Sometimes, if the task is rather challenging due to their language level or there is need for additional training, they might be asked first to explore another feature of the COCA corpus, e.g. Word in Context, i.e. they find on the display this feature and can query for immediate context of the given word with a certain number of words before and after, appearing in line. Thus, they can observe words and phrases in natural context, all texts in the corpus being authentic as known. It is a practical working out for students in generating authentic sentences afterwards when they have had exposure to authentic language.

Another variation of this task is observing, for instance, verb collocates, in real contexts, noticing tense usage, making frequency inferences. Also, it is useful to look for key words in different tenses, gather examples from the corpus samples and compile individual grammar glossaries. Students compare their results, analyze how different corpora represent the same grammatical categories, etc. Finally, they can proceed to using verb collocates in different tenses in sentences of their own.

- *Lemma versus word queries in Sketch Engine concordances.* Let's consider another example of corpus tools in the ESAP classroom. Concordances are a very efficient tool to find words and phrases in their natural context. For this purpose, I chose the Sketch Engine concordance query system which is a software program that can be used with beginner students (Nation, 2001). Students can either upload it onto their computers or work online. They will first learn how to use the simple search on the basic tab. They log in and switch into the new interface, then select a corpus in the required language, i.e. English, (I usually suggest them to choose a corpus of medium size because of speed limits) and then we proceed to basic search. They select Concordance, and after that they have two choices: *Lemma* or *Word.* In case they type "lemma", that is the basic form of the word, the search will automatically include all forms of the word as shown in the table below. Take for example, the lemma "Psychology", the search will give the following hits:

9352	K5M 5981	not accept the accuracy or validity of	psychologists	' findings. Their research, he said,
9354	K5M 6681	The hearing was told Smith was	psychologically	terrorised by her husband. She was
9355	K5M 8833	coding COLOUR, as an antidote to	psychological	and circumstantial greyness, is used
9357	K5M 9132	as coffee, which makes	psychologists	think it is the caffeine in both which
9358	K5M 10533	a dumping ground for	psychotics	and maniacs. I can't see how they achieve
9361	K8U 1562	faculty … a very real war	psychosis	' (quoted Posner, 1970, p. 64).
9363	K8V 2630	plus torture. So you were looking for a	psycho	A bit of a sadist, maybe with a touch
9369	K8Y 1562	could also be due to the greater	psychological	affinity between Is. In the case of
9370	K8Y 1644	Furthermore, illnesses of	psychosomatic	nature can be 'cured' by any
9371	K8Y 1661	, be designed to eliminate	psychological	biases in patient and physician.

Table 5: *Lemmas for 'Psychology' in the COCA Corpus*

Another type of activity with lemma search results could be the gap-filling exercise, performed in paper-based or online forms. In this exercise, students will infer the meaning of missing word, they will deduce which part of speech it is, the grammatical categories it has, etc. It is true that teachers can use any texts for this purpose, but it is proved that a simple text will not provide sufficient clues to infer the meaning. Corpus samples are rich in examples that elicit the meanings of the words, displaying a large number of contexts.

9352	K5M 5981	not accept the accuracy or validity of	………….	' findings. Their research, he said,
9354	K5M 6681	The hearing was told Smith was	………….	terrorised by her husband. She was

9355	K5M 8833	coding COLOUR, as an antidote to	…………	and circumstantial greyness, is used
9357	K5M 9132	as coffee, which makes	…………	think it is the caffeine in both which
9358	K5M 10533	a dumping ground for	…………	and maniacs. I can't see how they achieve
9361	K8U 1562	faculty … a very real war	…………	' (quoted Posner, 1970, p. 64).
9363	K8V 2630	plus torture. So you were looking for a	…………	A bit of a sadist, maybe with a touch
9369	K8Y 1562	could also be due to the greater	…………	affinity between Is. In the case of
9370	K8Y 1644	Furthermore, illnesses of	…………	nature can be 'cured' by any
9371	K8Y 1661	be designed to eliminate	…………	biases in patient and physician.

Table 6: *Gap-filling Activity with Sample from COCA Corpus*

In corpus search, students very often make a query for word, not lemma. How is this option different from the previous one? The learners will have hits for the word form "Psychology", they will automatically get only the word itself in the hits, and other forms of the word will be dropped out. For this kind search, there is a variation as well. Students will make out the meaning of the word. They also make decisions if the word has exactly the same meaning in all contexts.

9356	K5M 9131	Survey published in the	Psychology	Journal. A cup of tea also has an effect, though not
9368	K8Y 1441	agriculture, biology, medicine and even	psychology	, formal laboratory or field experiments can usually be done
9372	K91 81	At least in his judgment of French national	psychology	, Falkenhayn's appreciation had been accurate. Now the 'bleeding

Table 7: Word Query for 'Psychology" in Coca Corpus

- *Corpus-Based Error Correction*. Error analysis became a method on its own since the 1970s due to the reassessment of the importance of errors in ESL and EFL learning. Thus, according to Corder (Corder, 1981, p. 81), learner's errors are not random but systematic, random errors occurring only in the native language. He claims that such errors are not negative or interfering with target language but represent an elicitor, a facilitator in learning the new language.

For the error analysis task, students will be given texts of 450-500 words in length for translation. On the teacher's part, the error analysis framework will include collecting data, identification and classification of the types of errors, frequency data and contribution to preventing errors in the future. On the students' part, they will analyze the highlighted mistakes and do the corrective work afterwards. It should be noted that translation was made without any sources whereas the correction of mistakes was accomplished by means of corpus software. For this task I used the English Learner Translation Corpus (ELTC). The main goal was to analyze linguistic errors, and to a lesser extent, the translation errors. Therefore, students picked up examples from sample sheets and corrected the mistakes. Apart from that, they were made to classify the types of errors (morphological, syntactical and collocational). Hence they fulfilled some tasks to combat mistakes and improve their skills.

Thus, data-driven learning, by means of corpus work, proves to have an important correctional function. Students compare their writing with native speakers or check with the learner corpora to find common errors. It is a method that is becoming more prevalent in EFL classrooms, and ESP classrooms in particular. The experiments carried out prove that this method can be used even at the beginner and pre-intermediate levels when teaching grammar and vocabulary. Students can learn how to make generalizations both from online sources and paper-based samples with equal efficiency.

To sum up, hopefully, the wide array of classroom activities presented above as well as the students' feedback were able to break the long-existing stereotypes about corpus tools in language teaching. In recent times, EFL teachers would think that corpus-based or corpus-driven teaching is too sophisticated, time-consuming and inappropriate for teaching environments because of the knowledge, skills and technology that they must use in the classroom. However, there is growing evidence that many teachers set off such initiatives with minimal computer resources. They can make use of corpora in the classroom in different forms: online corpora tools, paper-based material for students to research, etc.

As Johns claims, the use of corpora in language teaching has been connected to a "data driven" approach, however we cannot claim that corpus use is restricted to any single teaching methodology. The use of corpora is attuned to all methodologies that accept explicit focus on language structure and use; in other words, teaching frameworks that lay stress on observation and awareness of language.

Corpus use can also enhance learner independence. According to Johns (Johns, 1991, p. 101), when using corpora or corpus-based materials, "students define their own tasks as they start noticing focus on language structure and use; in other words, teaching frameworks that reserve a role for noticing or awareness/consciousness-raising. As a final point, it is an invaluable tool which follows two-fold purposes: students embrace technology and use it in the classroom to increase authentic language awareness through authentic native speakers samples of text in well-balanced representative corpora.

References

Aarts, B. (2000). Corpus Linguistics, Chomsky and Fuzzy Tree Fragments. In *Corpus Linguistics and Linguistic Theory*. Ch. Mair, M. Hundt (Ed.).

Biber, D. (1993). Representativeness in Corpus Design. In *Literary and Linguistic Computing*, 243-257.

Biber, D. (2001). *Using Register-Diversified Corpora for General Language Studies*. Northern Arizona University Press.

Corder, S.P. (1981). *Error Analysis and Interlanguage*. Oxford University Press.

Gabrielatos, C. (2005). "Corpora and Language Teaching: just a Fling or Wedding Bells?" In *TESL-EJ*, 8(4), 1-35. http://www-writing.berkeley.edu/TESL-EJ/ej32/a1.html.

Johns, T. (1991). *Should You Be Persuaded: Two Samples of Data-driven Learning Materials*.

Nation, I.S.P. (2001). *Learning Vocabulary in Another Language*. Cambridge University Press.

EJ, 8(4), 1-35. http://www-writing.berkeley.edu/TESL-EJ/ej32/a1.html.

Rowley, J. (2007). The Wisdom Hierarchy: Representations of the DIKW Hierarchy. In *Journal of Information and Communication Science*. 10.1177/0165551506070706. S2CID 17000089.

Thomas, J. (2016). *Discovering English with Sketch Engine*. Versatile.

Corpus of Contemporary Americaon English. https://www.english-corpora.org/coca/?b=x2&c=coca&q=22285649.

Bun de tipar 01.12.2017. Garnitura Book Antigua. Comanda nr. 167. Tiraj 100.
Tipografia Universității de Stat „Alecu Russo" din Bălți. Mun. Bălți, str. Puşkin, 38.