CZU 811.111(072.8)

FIRST UNDERTAKINGS IN BUILDING A SPECIALIZED CORPUS BY SOCIAL ASSISTANCE STUDENTS

Liliana COŞULEAN, *University Assistant,*
Faculty of Philology,
Alecu Russo Balti State University

Rezumat: *În prezentul articol voi împărtăși propria mea experiență de compilare a corpusului, pe care am avut-o cu studenții anului I de la specialitatea de asistentă socială. Voi explica pas cu pas modalitatea în care elevii au colectat texte, au compilat corpusuri și, în cele din urmă, cum au trebuit să extragă informații utile din propriile corpusuri, îndeplinind diverse sarcini sugerate de profesor. Există o mare varietate de activități didactice, care pot fi realizate în cadrul unui corpus specializat. Scopul articolului a fost să demonstreze, că principalele caracteristici ale corpusului - listele de frecvență și șirurile de concordanță - sunt foarte utile la învățarea englezei profesionale și au devenit instrumente indispensabile în predarea limbajului autentic.*

Cuvinte-cheie: *corpus software, liste de frecvență, concordanțe, corpusuri, corpusuri paralele, variație de registru, corpusuri balansate, corpusuri specializate.*

From the very start, I would like to make a short introduction to what corpus linguistics is, and eventually how it can be related to English language teaching. There is a very broad and simplistic definition of Corpus Linguistics as a regular tool/ method in language teaching. The most prominent scholars in corpus linguistics such as G. Leech, D. Biber, S. Conrad, R. Reppen, M. McCarthy, V. Brezina, J. McEnery, and of course J. Sinclair argue that it is not just an approach, a method, an instrument, but first and foremost an ample study of language using corpora (plural for corpus). [1, pp. 23-24]. A more specific definition of a corpus is a principled and large collection of authentic texts, i.e. a body of texts that are stored electronically, in a computer, and analyzed using special software designed for corpus analysis. In a nutshell, the scope of corpus is as follows: when a researcher is interested in a certain linguistic phenomenon, they will collect specific data (corpus) depending on research objectives and analyze them using this specific software installed on their computer. [1, pp. 16].

In the present article I would like to share my own corpus compiling experience that I had with my university students. I will explain step by step how students collected texts, compiled corpora and ultimately how they had to extract useful information from their own corpora, fulfilling various tasks suggested by the teacher.

Before starting to download texts, students should set their research objectives. It is necessary for the eventual corpus design that texts suit the research goals. Subsequently, students should decide where from to collect their data in such a way that authorship issues should not appear.

So where to collect the data from? Should students go outside and interview people? Should they listen to broadcast conversations? For ESAP learners, I suggest to get them compile their own specialized mini-corpora using authentic texts on social assistance issues published on news portals in the internet. Therefore, students had to choose the topic-based articles from the news sites like BBC, New York Times, CNN, Washington Post, the Guardian, the New York Times news, etc. to be noted that we made specific requirements for the British or American variety of English.

The next step in building the corpus consisted in having special software downloaded onto the students computers. There are various softs available that come in handy for a language instructor, or a researcher, in broader contexts, in order to carry out language teaching or more serious linguistic investigation. Nowadays the most popular last generation tools in corpus methods of teaching are considered LancsBox and Antconc. Users prefer them to many other software due to the multitude of useful functions displayed and additional advantages such as being free, taking small space on the computer and being user-friendly. LancsBox is a new generation corpus-analysis tool designed by V. Brezina, M. Timperley, D. Gablasova and T. McEnery. The software is designed for researchers, students and teachers. The other software AntConc was designed by L. Anthony, also of Lancaster University. The AntConc version 3.5.8 is a last generation tool, easy to use and provides hits at high speed and great query results.

Another issue we have to deal with is the best corpus design. During the process of copying and downloading the texts from news portals, students must take decisions concerning formatting since the files will be saved in electronic form. If they use for instance AntConc , then files can be saved in a Word, a PDF or plain FTC format.[10] As far as I know, LancsBox reads PDF files. So, you can simply import them there and work on them without the need to convert them to .TXT files. If you want to use AntConc however, then I'm afraid you'll need to convert all the documents to .TXT files.

At the stage of selecting texts from the WWW, there is a little confusion about copying texts and uploading them from the news site onto the computer. When you click on an article to select it, then other annoying texts or images will get selected as well. It is common experience which is not pleasant at all. To avoid such situations, there is a special tool JusText demo, which has the role to clean the web-sites and we can copy just plain texts, without headers, links, images, advertisements, etc. Therefore, we copy the article link, paste it into the URL box in the JusText demo tool,[11] click SEND, and get our article clean. In this layout we can copy it and save it for our corpus in the UTF-8 format as this encoding reads all languages. Depending on how large they want their corpus, students

[10] https://www.laurenceanthony.net/software/antconc/

[11] https://pypi.org/project/jusText/very

can make this operation at least 20 to 30 or more times, thus saving about 30 news articles and collecting a corpus of about a couple of thousand words.

Finally, the students' corpora are ready and they can already start working on their own corpus. They can load the files and have them on the AntConc tool.

As a demo, I would like to present some results of their research into the corpus. The phrases analyzed are ***social assistance, social care*** and ***social welfare.***

Frequency lists. One of the first tasks that they can try on this application is to generate a frequency list. Since corpus linguistics has so much to deal with frequency, students will explore the quantitative characteristics of the corpus. [2, p.14] With the frequency list feature, they will observe words from the most frequently used one to the least frequent and draw conclusions. For this purpose, they will go to the word list tool, click on key word and they will see a list of most frequent words in their corpus. They see information about how many times this word is used in the corpus. Also, they see how many tokens their corpus is made of. Students will see that the most frequent words in the list are function words; they are linking verbs, prepositions, conjunctions. Content words come at about line 17 and below in the frequency list in the COCA. An excellent activity for students will be to fill in the chart with the missing social assistance terms that they will find in their corpora. They will also write the line number, i.e. the frequency order in which they appear within the corpus. It is a useful exercise for comparing different result for most used social assistance terms.

Line No	General English / function words	Line No	Social Welfare English
1	**one**	?	
2	**said**	?	
3	**No**	?	
4	**Time**	?	
5	**Only**	?	
6	**Other**	?	
7	**Two**	?	
8	**Now**	?	
9	**Very**	?	
10	**First**	?	

Table 1. *Function words versus content words activity in the COCA frequency list*

Another example of word frequency is exploring the Google engine for certain terms frequency of occurrence. Thus, students will type in the phrases "social assistance", "social welfare" and "social care" in regular Google search and Google Scholar and see what frequency results it they show. To compare the results, they will look for the same phrases in a corpus at their choice. Later they will compare the results and make assumptions. You can see below what hits are given by the two engines and one randomly chosen corpus.

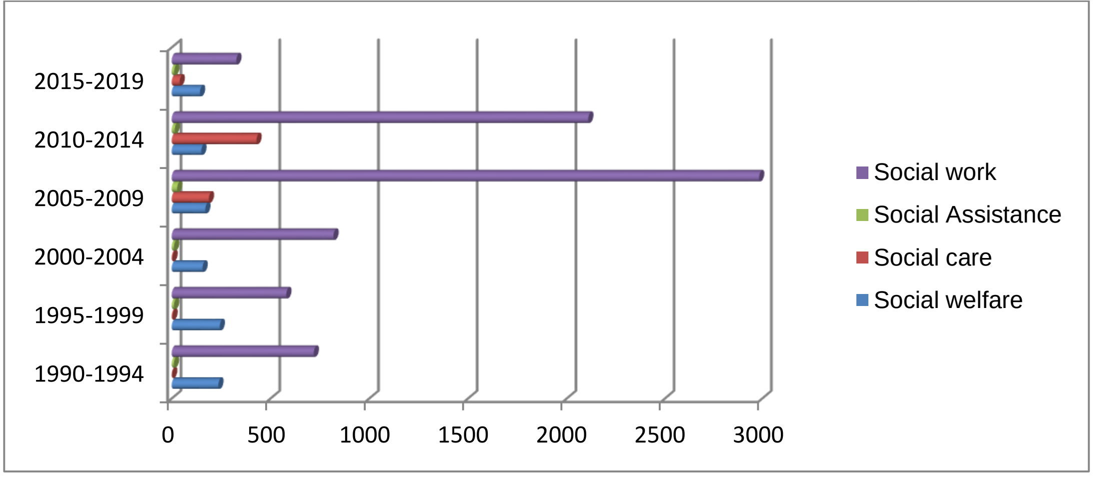
Term	Google	Google Scholar	BNC
Social assistance	8,270,000	180, 000	16,000
Social care	52,200,000	1,010, 000	198,000
Social welfare	44,200,000	2 540, 000	348,000
Social work	112,000,000	2,420, 000	1, 232,000

Table 2. *Frequency hits across Google, Google Scholar and British National Corpus*

An example of contrastive research that students can carry out is compare frequency of words or lemmas from the point of view of **gender**. In this respect, a very interesting research was conducted by Vaclav Brezina, a renowned corpus linguist from Lancaster University. He was interested in what can corpora tell linguists about learning a foreign language. Quite a lot. For this reason he analyzed a large corpus of texts written by men and women and discovered a considerable difference in frequency of use of pronouns. As a result, he concluded that female British speakers use pronouns more frequently than male British speakers. Definitely, the computer cannot tell the reason why it happens; this is one of its limitations. Nonetheless, such an exercise would be beneficial to students in making inferences about language use based on gender, social context, age, etc., speculating and expressing their own point of view on linguistic and sociolinguistic phenomena. Applied directly to their own corpora, students will search for certain words or lemmas and compare results for male and female texts.

There is a wide variety of similar teaching activities that can be done within a corpus [5, pp. 37-62]. Related to the social assistance, another useful activity is that of comparing synonymous terms diachronically. For this reason, let us take the following phrases ***social welfare, social assistance*** and ***social care***. What information can students get from the corpus tools query? We shall analyze them in the COCA[12] corpus using the diachronic search feature within the frequency line option. It is possible to do that because the corpus contains texts beginning with 1990s till nowadays. The results may surprise our students. They will find out that these synonymous terms had high variability of usage throughout certain periods of time that differ greatly from one another. It comes out that **social work** used to be much more popular than **social care** at a definite time in the past, while **social welfare** emerged at later periods. Students are encouraged to make tables, diagrams and compare data in order to make conclusions about diachronic development of a word in terms of usage preferences.

Table 3. *Diachronic search results in the COCA corpus (based on frequency lists)*



The activities above make part of a far larger methodology which involves corpus exploration and work with distinct features of corpus software, especially designed for teaching and learning activities. As it could be noted, for a language instructor, learner corpora are endless resources of genuine, authentic material that can be explored and used in many ways both by language teachers and learners.

Bibligoraphy:

1. BIBER, D., CONRAD S., REPPEN R., *Corpus Linguistics: Investigating Language Structure and Use* (Cambridge Approaches to Linguistics) 1st Edition, 1998, 288 pp., **ISBN 0-521-49-622 -5**
2. GRIES, S. T. *Quantitative corpus linguistics with R: A practical introduction*, 281 pp, 2016, **ISBN 978-1-138-81628-2**
3. McENERY, T., & HARDIE, A., *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, 2012, 294 pp, **ISBN 978-1-13-85981-4**
4. SOEHN, J. P., ZINSMEISTER, H., & REHM, G. *Requirements of a user-friendly, general-purpose corpus query interface*, 2008
5. TRIBBLE, C. (2015). *Teaching and language corpora: Perspectives from a personal journey*. In Leńko-Szymańska, A. & Boulton A. (Eds.) Multiple Affordances of Language Corpora for Data-driven learning.312 pp, **ISBN 978-9-02-72687-16**
6. **https://www.english-corpora.org/coca/**

[12] https://www.english-corpora.org/coca/