ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

Analysing Some Uzbek Texts Via Corpus Analysis Toolkit - "Antconc"

Tangriyev Valisher Azmovich

Teacher of the faculty of Foreign philology, Termez State University, Uzbekistan

valishert@mail.ru

Abstract: *This paper gives some information about Corpus linguistics and its prospects. We are using corpus analysis software called AntConc to process the corpus data. In general, this program is used to analyze text to find a certain pattern in language. Corpus linguistics is a study of language that includes all processes related to processing, usage, and analysis of written or spoken machine-readable corpora. This report will explain the steps of processing the data with this software. We analyzed the famous Uzbek writer Said Ahmad's three stories "Sarob", "Qishdan qolgan qarg'alar" and "Azroil" with the help of free software AntConc.*

Key words: *Corpus, Corpus linguistics, AntConc, Concordance, Word List, Concordance Plot, File View.*

There are many "hyphenated branches" of linguistics, where the first part of the name tells you what particular aspect of language is under study: sociolinguistics (the relation between language and society), psycholinguistics (the relation between language and the mind), neurolinguistics (the relation between language and neurological processes in the brain) and so on. Corpus linguistics is thus a methodology, comprising a large number of related methods which can be used by scholars of many different theoretical leanings. On the other hand, it cannot be denied that corpus linguistics is also frequently associated with a certain outlook on language. At the centre of this outlook is that the rules of language are usage-based and that changes occur when speakers use language to communicate with each other [Hans Lindquist 2009].

We are at the boundary of a new era of English language studies. The scientific study of the language had just two broad traditions, until the last quarter of the 20th century. The first was a tradition of historical enquiry, encouraged by the comparative studies of the 19th – century philology. The second was a tradition of research into the modern language, fostered chiefly by the theories and methods of descriptive linguistics. Now there is a third attitude to consider, stemming from the results of the technological revolution, which is likely to have considerable effects on the goals and methods of English language research. Because of the

 Copyright ⓒ 2019Authors

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

great availability of computers and its processing of natural language has made it able to achieve to collect huge amounts of data – corpora were developed.

According to David Crystal "Corpora are large and systematic enterprises: whole texts or whole sections of text are included, such as conversations, magazine articles, newspapers, lectures, broadcasts and chapters of novels" [Crystal 2003].

There is no doubt that advances in electronic instrumentation and computer science have changed our perception of language. Corpus Linguistics has evolved due to the great opportunities offered by computer technology. According to the computational linguist and lexicographer John Sinclair: "**Corpus Linguistics is a study of language that includes all processes related to processing, usage, and analysis of written or spoken machine-readable corpora**" [Meyer 2002].

Corpus linguistics considers not only speech, but language as well, arguably, in a new light, providing a new vision and a more powerful potential, which derives from a mere quantity of language samples, a growing scale of their coverage [Gvishiani 2008].

Most lexical corpora today are POS-tagged. However even corpus linguists who work with 'unannotated plain text' inevitably apply some method to isolate terms that they are interested in from surrounding words. In such situations annotation and abstraction are combined in a lexical search [Johanson 1991].

The advantage of publishing an indexed corpus is that other users can then carry experiments on the corpus. Researchers and linguists who have done research in other fields can use this resource. By sharing data, corpus linguists are able to look at the corpus as a center of linguistic discussion, not a complete source of knowledge.

In recent years, corpus linguistics has experienced a great awakening. From being marginalized approach which is used mainly in English linguistics, and more specifically, English grammar, corpus linguistics has started its opportunities. The history of corpus linguistics exists almost as a component of academic folklore. It is not a new idea the use of collections of text in language study. Making lists of all the words in a particular text with their contexts began in the Middle Ages. Today

 Copyright ⓒ 2019Authors

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

we call it *concordancing*. Other scholars have calculated the frequency of words from single texts or set of texts and compiled a list of the most common words. Corpora used in areas of language include syntax, semantics and comparative linguistics. Even if the term "corpus linguistics" was not used, most of the work was similar to corpus-based research linguists.

A corpus and its processing are also virtually useless if there are no computer software tools for identifying the results. In this area there are two of the most general software tools: MonoConc Pro[1] , and WordSmith Tools[2] . Despite the fact that many other concordancers and corpus analysis programs have also been developed. A small number of these programs were designed in a special classroom environment for students. In this article, I will describe AntConc[3] , a corpus analysis toolkit designed specifically for use in the classroom. AntConc is a free program that is suitable for people with limited budgets, schools or colleges, and works on Windows and Linux / Unix-based systems. Despite being a free software license, it has a user-friendly, intuitive graphical user interface, and a powerful console, word and key frequency generators, cluster and lexical analysis tools, and so on.

How to use AntConc software

At the beginning of process, we must download the freeware AntConc from the internet, and then open the program to start using it. On the left-hand side, there is a window to see all corpus files loaded[4]. There are 7 tabs across the top:

- *Concordance: This will show you what's known as a Keyword in Context view (abbreviated KWIC, more on this in a minute), using the search bar below it.*
- *Concordance Plot: This will show you a very simple visualization of your KWIC search, where each instance will be represented as a little black line from beginning to end of each file containing the search term.*

[1] Information and download instructions available at: http://www.monoconc.com/
[2] Information and download instructions available at: http://www.lexically.net/wordsmith/
[3] Information and download instructions available at: http://www.antlab.sci.waseda.ac.jp/
[4] https://programminghistorian.org/en/lessons/corpus-analysis-with-antconc

Copyright ⓒ 2019Authors

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

- *File View: This will show you a full file view for larger context of a result.*
- *Clusters: This view shows you words which very frequently appear together.*
- *Collocates: Clusters show us words which definitely appear together in a corpus; collocates show words which are statistically likely to appear together.*
- *Word list: All the words in your corpus.*
- *Keyword List: This will show comparisons between two corpora. After the software is opened, we must upload the data that we want to analyze.*

Data must be checked, corrected converted *text file* or *txt file*, since the program only can read these types of files. After getting the data we can upload it by clicking file on the menu bar, after this we should click *open file*. By getting the data uploaded on this software, we can process it as we want *(figure 1)*. For discussion we have chosen the famous Uzbek writer Said Ahmad's (Saidahmad Husanho'jaev) three stories namely "Sarob" ("Mirage"), "Qishdan qolgan qarg'alar" ("The crows which have stayed since winter"), "Azroil" ("Azrael").

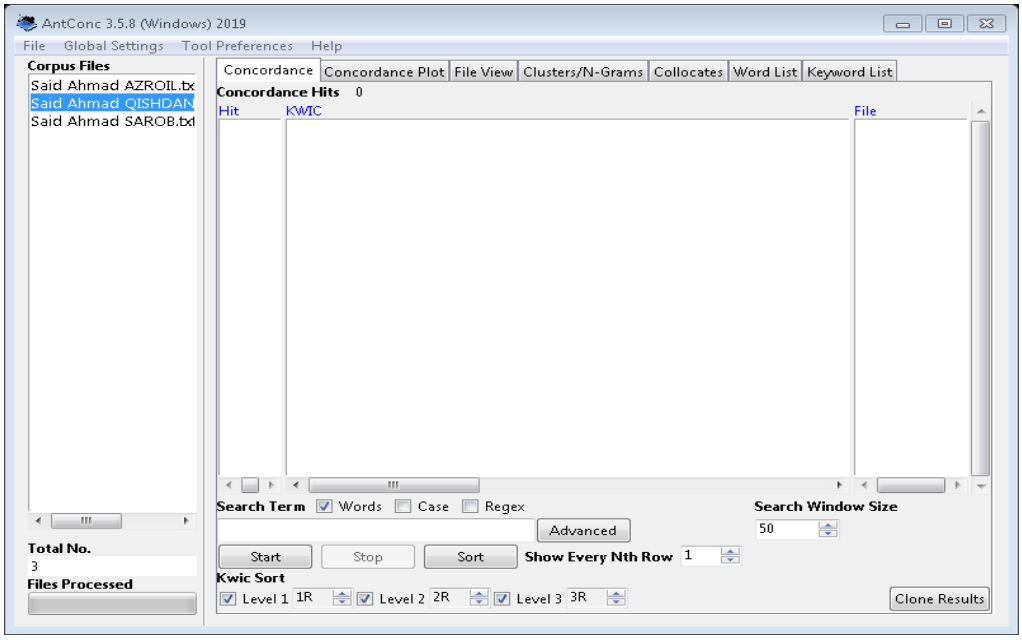


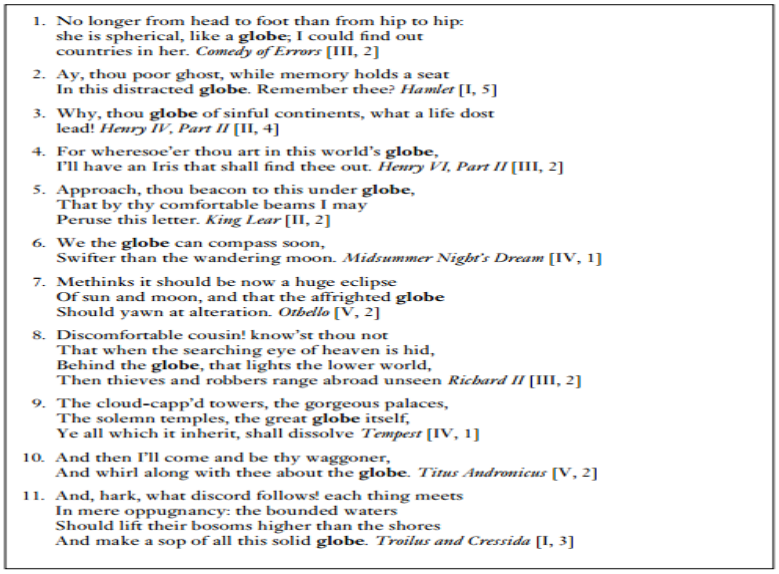
Figure 1. Uploading text or txt files to AntConc

Concordances

Copyright ⓒ 2019Authors

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

Traditionally, concordances are the lists of all the contexts in which a particular text contains a search term. For example, theologians and literary people gathered a collection of comments to be able to learn how to use certain words in the Bible and Shakespeare's work. For instance, if you look up the word "globe" in a Shakespeare's concordance, you will find out that the poet used this word 11 times. [Hans Lindquist 2009]. These can be displayed with the immediate context as in *Figure 2.*



1. No longer from head to foot than from hip to hip:
she is spherical, like a **globe**; I could find out
countries in her. *Comedy of Errors* [III, 2]
2. Ay, thou poor ghost, while memory holds a seat
In this distracted **globe**. Remember thee? *Hamlet* [I, 5]
3. Why, thou **globe** of sinful continents, what a life dost
lead! *Henry IV, Part II* [II, 4]
4. For wheresoe'er thou art in this world's **globe**,
I'll have an Iris that shall find thee out. *Henry VI, Part II* [III, 2]
5. Approach, thou beacon to this under **globe**,
That by thy comfortable beams I may
Peruse this letter. *King Lear* [II, 2]
6. We the **globe** can compass soon,
Swifter than the wandering moon. *Midsummer Night's Dream* [IV, 1]
7. Methinks it should be now a huge eclipse
Of sun and moon, and that the affrighted **globe**
Should yawn at alteration. *Othello* [V, 2]
8. Discomfortable cousin! know'st thou not
That when the searching eye of heaven is hid,
Behind the **globe**, that lights the lower world,
Then thieves and robbers range abroad unseen *Richard II* [III, 2]
9. The cloud-capp'd towers, the gorgeous palaces,
The solemn temples, the great **globe** itself,
Ye all which it inherit, shall dissolve *Tempest* [IV, 1]
10. And then I'll come and be thy waggoner,
And whirl along with thee about the **globe**. *Titus Andronicus* [V, 2]
11. And, hark, what discord follows! each thing meets
In mere oppugnancy: the bounded waters
Should lift their bosoms higher than the shores
And make a sop of all this solid **globe**. *Troilus and Cressida* [I, 3]

Figure 2. Concordance for globe in Shakespeare's work.

Source: Based on http://www.opensourceshakespeare.org

In most of the corpus analysis software's central tool, including AntConc, is concordance. As Sun & Wang [2003] describe, concordancers have been shown to be an effective aid in the acquisition of a second or foreign language, facilitating the learning of vocabulary, collocations, grammar and writing styles. Figure 3 shows a screenshot of AntConc while a user is operating the Concordancer Tool. In this screenshot we searched the word "kerak" (English equivalent should, need, mast). In the figure 3 we can see there are some symbols like "\x91", "\x97". These characters are the equivalent of Uzbek symbol and punctuation mark (ʻ) and (-) respectively. We presented these errors so that while preparing data one must pay attention this kind of pronunciation misunderstandings.

 Copyright ⓒ 2019Authors

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

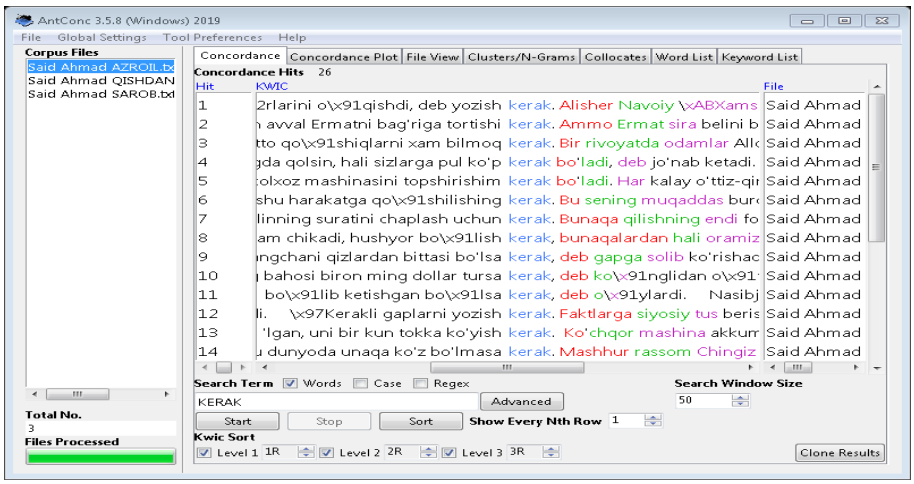


Figure 3. Concordance of the word "kerak"

Concordance tool is also handy who wants to learn word formation or collocations of any searched word. Like other tools in the software, Concordancer Tool gives you access to the most common operations directly on the home screen. Lonfils & VanParys [2001] explain that this is an essential feature of good software design as it avoids the need for confusing pull-down menus and additional windows.

This menu is used to show a list of phrases that contain the words we are looking for. We can search for any word by typing it in the search box at the bottom, and after clicking on the "start" button, a list of phrases with the highlighted word appears.

Above the searching box there are three main functions; they are *words, case and regex.* If we checklist the *word* function, then we enter the word "pul" for example, then there will be list of sentences which contain the word "pul". However, if we do not checklist the *word* function then we can find the words "pul", "pul berish", "pul bilan", and "pul bor" and so on. Then the *case* function is used to show the list of the sentences which contain word in certain case. For example, if we checklist the *case* and we enter the word "Alloh" in uppercase then it will show only the sentences which contain the word "Alloh" in uppercase.

In addition, we also can find the word variation of *odam*, like "odamga", "odammas", and "odamlar" by entering *odam* in the searching box. It will show

 Copyright ⓒ 2019Authors

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

every word that contain *odam* and the words are all in highlight. It is different with the *word* function because it only highlight the *odam* not the others.

Based on the result, we find many variations of "odam" like "odamga","odammas", "odamlar", and so on. We can enter the word one by one to see the frequency of each word. For example, we can enter "odamlar", and see how many times it appears. We can see the number of its appearance in *Concordance Hits*, and it shows that "odamlar" appear for 13 times in the three mentioned stories. Then "odamga" appears 7 times, and "odammas" appears once only.

Word List

When analyzing corpus data, we firstly need to know how many words, what words, and how often the words appear in the text. We can use *word list* function to get that data which is located on the menu bar. After uploading the data, we can go to the *word list* menu, and then click "*start" button* that is located on the bottom. Thereafter, we can see the words used on the text that is showed in series based on what words mostly used. In this menu, we also can meet the *word types* and *word tokens.* To take the example, if we upload the *txt file* of Said Ahmad's three stories "Sarob" ("Mirage"), "Qishdan qolgan qarg'alar" ("The crows which have stayed since winter"), "Azroil" ("Azrael"), we can find the amount of words used in the text on *Word Types (Figure 4)* that are 6716 words, with the word "bir" in the sixth rank as the highest rank because it's the most frequently used word on the text (**Note**: At the beginning of the table 5 characters which is given in the rows from 1 to 5 have no any meaning in Uzbek language. They are misunderstandings of the pronunciation marks and symbols which are mentioned before. "x" is taken from "\x91", "\x97", "o" is the beginning of the Uzbek letter(o') for example, "o'chdi", "o'g'li", "o'jar", "o'ksib-o'ksib" etc., "bo" is the beginning of the words like: "bo'ladi", "bo'g'iq", "bo'lib", "bo'rtib" etc., "ko" is the beginning of the words such as: "ko'cha", "ko'hlik", "ko'kka" etc., "qo" is the beginning of the words like: "qo'l", "qo'chqor", "qo'lansa"etc.). We also can see the amount of *word tokens* or the total words that is 19499 words.

 Copyright ⓒ 2019Authors

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

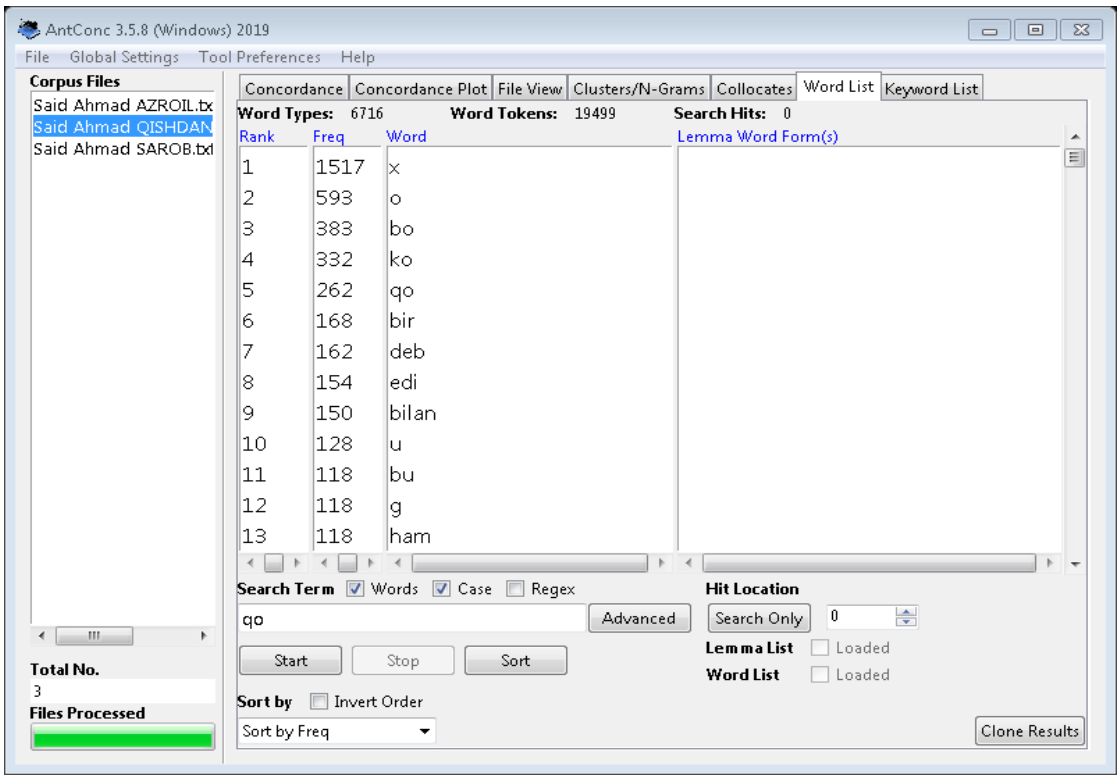


Figure 4. Word list of Said Ahmad's three stories

If we observe the 20 most used words, we can see that most of the words used in the text are belongs to different part-of-speech. There is only one word which is included as number or pronoun on the number 10, 11, and 18. Therefore, based on the data, we can conclude that people actually used different types of part-of-speech.

In this menu, we can also calculate the token ratio by calculating the formula; it is used to measure the diversity or richness of words that people have, with formula *word types* /*word tokens* x 100. For example, based on the same file the type token ratio is 6716/19499 x 100 = 34.44%.

Concordance Plot

The appearance of the word we are searching for is shown in this menu. For instance, on the same file when we want to know how the word "*odam*" is placed, we can see the sketch of the word and the number of its occurrence. Based on the following figure (*figure 5*), it can be seen that the word "odam" appears for 39 times and it emerges different parts of the text.

 Copyright ⓒ 2019Authors

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

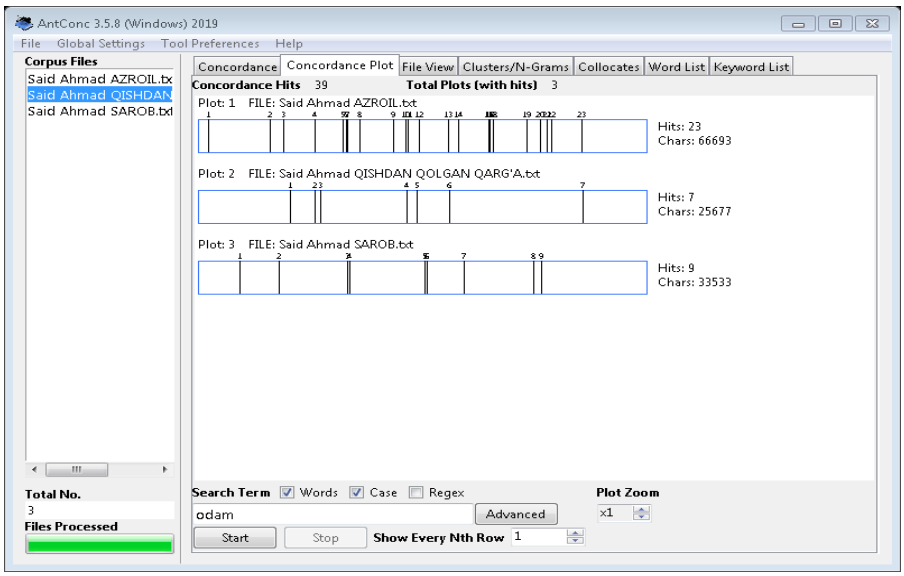


Figure 5. Concordance plot of the word "odam"

File View

If we want to see the whole text, through this menu we can show the entire text and the word we are looking up. To give an example, if we enter the word *"edi"* in the searching box of *file view*, we can see all sentences and the *"edi"* are in highlight to make the reader absorb it with no difficulty (*Figure 6*).

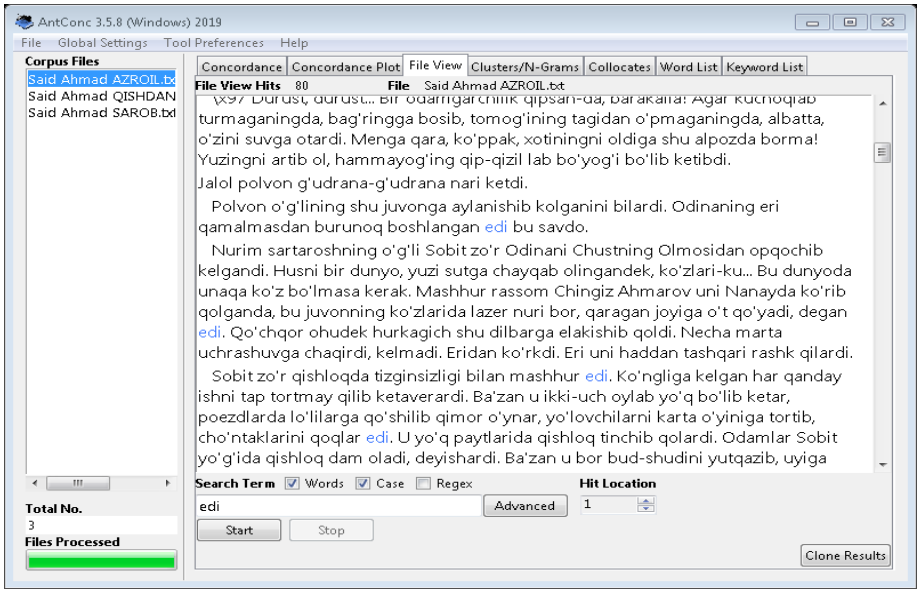


Figure 6. File view of the word "edi"

Conclusion

To sum up, corpus linguistics has become a rapidly developing field of world linguistics, and the corpus itself is an invaluable tool for linguists in the age of

 Copyright ⓒ 2019Authors

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

information. Corpus linguistics, as a whole, ensures full coverage of all aspects of language phenomena, guaranteeing the originality of information. Corpus analysis is a form of text analysis that allows for large-scale comparisons between text objects. As people read, we can see things that we absolutely do not see. If we have got a collection of documents, we may want to find patterns of grammatical use, or frequently recurring phrases in our corpus. We also may want to find statistically likely and/or unlikely phrases for a particular author. Corpus analysis is especially useful for testing intuitions about texts and/or triangulating results from other digital methods[5].

By working with this freeware AntConc we have learnt:

- create/download a corpus of texts
- conduct a keyword-in-context search
- identify patterns surrounding a particular word
- use more specific search queries
- look at statistically significant differences between corpora

There are seven tabs in the AntConc menu, but in this paper we presented three important function of software: Concordance, Word List, Concordance Plot and File View. Due to the given stories are not annotated or there is no any Uzbek Pos-tagger software we have not mentioned other functions of AntConc.

Used literatures:

1. Алякринский А. П. “Investigation of Semantically-Related Words “Small/Little” in the British National Corpus” pp. 1-8.

2. Гвишиани Н. Б., English on Computer. A Tutorial in Corpus Lingustics. - Москва: Высш. шк., 2008. pp. 191-192

3. Crystal D. The Cambridge Encyclopedia of the English Language. - Cambridge University Press, 2003.

4. Hans L. “Corpus Linguistics and the Description of English” Edinburgh University Press 2009 p.1-8.

[5] https://programminghistorian.org/en/lessons/corpus-analysis-with-antconc

 Copyright ⓒ 2019Authors

ISSN:0971-1260
Vol-22-Issue-4-October-December-2019

5. Johanson S., Stenstrom A. “English Computer Corpora: Selected Papers and Research Guide”. - Walter de Gruyter, 1991. pp.127-148

6. Lonfils, C. and Vanparys, J. 2001. How to design user-friendly CALL interfaces. Computer Assisted Language Learning, 14(5), 405-417.

7. McEnery, T; Wilson A; “Corpus Linguistics”. Edinburgh University Press Ltd.2005. pp. 3-22

8. Meyer Ch. “English Corpus Linguistics: An Introduction”. - Cambridge University Press, 2002. pp. xvi - 168

9. Said A. “Tanlangan asarlar” III jildlik-I jild “Hikoyalar”, “Sharq” nashriyot-matbaa konserni bosh tahririyati Toshkent – 2000 pp.7-39

10. Sun, Y. C. & Wang, L. Y. 2003. Concordancers in the EFL Classroom: Cognitive Approaches and Collocation Difficulty. Computer Assisted Language Learning, 16 (1), p. 83-94

Websites:

1. http://www.monoconc.com/
2. http://www.lexically.net/wordsmith/
3. http://www.antlab.sci.waseda.ac.jp/
4. https://programminghistorian.org/en/lessons/corpus-analysis-with-antconc

Copyright ⓒ 2019Authors