

Compiling Dictionaries by Using Corpus Analysis and its Advantages

Nozimjon Ataboev Bobojon O'g'li

PhD student at Uzbekistan State World Languages University



Abstract - As we all know about the role of computers in our present-day life, it should be taken into consideration that all spheres of the scientific investigations are being interfered by this very technology. Especially, its contribution to the linguistics has been considerably substantial. Since, the computational linguistics, that is, the branch of linguistics in which the techniques of computer science are applied to the analysis and synthesis of language and speech, has already become one of ways to find out possible solutions to the topics under hot discussions and this linguistic field is popularizing the value of computers in using language investigations. Here, one needs to keep in mind that the development of the computer linguistics was also essential while speaking of the role of computer aided spheres. Historically, this whole aggregation of the scientific achievements has positively influenced on the emergence of the new linguistic branch named *corpus linguistics*. In the following article we will try to focus on informing about this newly-emerged science and its application in lexicography.

Keyword - *Corpus Linguistics; Corpora; Corpus Analysis; Lexicography; Dictionary-Compiling; Word Frequency.*

The fast-growing demand and interest of practitioners and researchers in the field of Corpus Linguistics, particularly in application of corpus analyses in different aspects of linguistics and teaching foreign languages has been emerged in linguistics.

Table 1. The fields available to be used from the corpus analysis

<i>lexicography</i>
<i>sociolinguistics</i>
<i>discourse analysis</i>
<i>morphology</i>
<i>phonology</i>
<i>syntax</i>
<i>teaching</i>

We might want to center around the application of corpora in arranging word references, i.e. "corpus-based lexicography". Before moving to the expression "corpus-based lexicography", it would be better to feature the historical backdrop of word reference making. The act of lexicon arranging started as right on time as 1600 when

Robert Cawdrey included words that were regarded troublesome as they were acquired from different dialects into his variant of the lexicon. Cawdrey included into his lexicon obtained words giving them through definitions, equivalent words, cases of utilization and collocations. Samuel Johnson presented the techniques for assembling word references in the 1700s which later were utilized to make "A New Dictionary" that is known as the Oxford English Dictionary. As is self-evident, gathering lexicons was exceptionally troublesome and tedious occupation, in light of the fact that with a specific end goal to decipher a word the lexicon producers needed to examine more than three or four printed books. That required a great deal of physical vitality also. Additionally, the instant lexicons were not completely conceivable for every one of the students and still are not a direct result of hard words and absence of assortments of choices and illustrations.

The advancement of Corpus Linguistics has brought forth *Corpus-based Lexicography* and another corpus-based age of word references. A vast and all around built corpus gives phenomenal data about recurrence, conveyance, and

normality of semantic highlights –, for example, words, collocations, spellings, articulations, and linguistic developments.

As the English researcher Meyer declares Corpus has given a great deal of commitments in dialect contemplate since its rise, however its effect on etymology began in 1989. Because of progress of PC programming, mechanized information word specialists would now be able to spare their chance and the enormous measure of work required for ordering a lexicon. Regularly, a lexicon incorporates data with respect to discourse, use, which means, elocution, historical underpinnings of a word. Prior to the coming of corpora, this data must be accumulated physically; word specialists expected to do the hard work of gathering pieces of paper containing content that they mean to incorporate into the lexicon. Therefore, it took approximately 50 years to finish Oxford English Dictionary, which was later known as New English Dictionary.

The upsides of utilizing corpora in word reference ordering incorporate the followings:

- a) with corpora, lexicon producers would now be able to utilize an expansive example of credible talked and composed content as a source to delineate how each word in their rundown is utilized as a part of genuine [2]. The reference utilized as a part of lexicon originates from genuine talk which gives exact, all around characterized lexical implications in the meaning of a word in word reference;
- b) one immense change in lexicon making is the rich data accessible for words that have numerous invariant implications, for example, 'take', 'go', and 'time', which have a tendency to be disregarded in the past word reference rehearse [3];
- c) another enormous preferred standpoint of utilizing corpora in etymology is that data on word recurrence can likewise be

acquired. Along these lines, etymologists can allocate whether a word is among the initial 500 most basic words, the following 500, et cetera [2]

For example, the British National Corpus (both composed and talked) has more than 100 million words. The composed section (90%) incorporates removes from daily papers, scholastic books, school and college articles, and so on. The talked section (10%) incorporates a lot of unscripted casual discussion, recorded by volunteers chose from various age, locale and social classes in a demographically adjusted manner. That sum empowers the etymologist to make an exceptionally insightful and counter choice on the words. In addition, this corpus speaks to the present-day English dialect, which makes it the wellspring of genuine cases.

That is the reason, one can without much of a stretch make conclusion that the utilization of corpora in lexicon making rehearses gives a word specialist a great deal of chances, among the most critical ones are:

- 1) to create and overhaul lexicons rapidly, in this manner giving up and coming data about the dialect;
- 2) to give more entire and exact definitions since a bigger number of common cases are analyzed;
- 3) to keep over new words entering the dialect, or existing words changing their implications;
- 4) to depict utilizations of words or expressions regular for specific assortments and classes;
- 5) to arrange illustrations separated from corpora with a specific end goal to clarify words and collocations;
- 6) to treat expressions and collocations all the more efficiently because of the presence of common data apparatuses which build up connection between co-happening words;
- 7) to enlist social undertones and hidden belief systems which a dialect has.

Table 2. A. Kilgarriff's point of view on the aspects of dictionary creation through corpus application

<i>headword list development</i>	<i>for writing individual entries</i>			
their syntactic behaviour, the collocations they participate in, any preferences they have for particular text-types or domains	discovering the word senses and other lexical units	identifying the salient features of each of these lexical units	providing examples	providing translations

Corpora can make dictionary compiling more exact, proficient, complete and reliable. They can convey a hopeful headword list, and, where the corpora are produced in light

of care with neologism-finding, can recognize competitor neologisms.

There are numerous manners by which they can bolster passage composing. They can give an extensive variety of pieces of information to the etymologist for breaking down the word's scope of importance into particular detects. In mix with an appropriate corpus inquiry framework they can discover the sayings, expressions and collocations for a word. They can distinguish if a word has important conduct in connection to syntax, space, locale and enlist. They can do the preliminary work for discovering great case sentences, and interpretations. Corpora have been utilized as a part of these courses in a scope of lexicon ventures, and the section has depicted how this has functioned, with reference to a specific corpus question device, the Sketch Engine. Throughout the most recent two decades, the word specialist's part has been increasingly regularly, checking and affirming or altering the corpus device's work, where prior it would have been 'composing starting with no outside help'.

Edita Hornáčková Klapicová, in her article 'Piece of the Entry in a Bilingual Dictionary', talks about the ordering procedure of bilingual lexicons and notices Zgusta's purpose of perspectives. Zgusta conceded that the different significance of the lexical unit of the source dialect is essentially the premise of the entire development of the bilingual section, of the arrangement of the faculties. Zgusta includes that separated from every one of these signs, the passage of a bilingual lexicon may give diverse other data associated with the coordination of the lexical units. This could be some all encompassing data, derivation of the lexical units or data on the lexicalized and the verbatim significance of various morphemic and word blends. Taking all things together, Zgusta's criteria for the development of the section in a bilingual lexicon could be condensed as following:

- a. The nearness of the section word in its sanctioned frame;
- b. Grammatical data;
- c. Indication of articulation;
- d. Equivalentents in the objective dialect in their accepted shape;
- e. Indication of the entire lexical importance of the passage word by fractional counterparts of the objective dialect;
- f. Encyclopedic data;
- g. Etymology of the passage words;
- h. The lexicalized and the verbatim importance of various morphemic and word blends.

Coming out from the criteria by Zgusta, we can expect that a huge corpus is the main answer for a large

portion of the checked necessities, these days. We consider that it is smarter to break down every rule one by one.

a. The use of the word in the lexicon is the least demanding activity like writing the word in the inquiry line;

b. A corpus can demonstrate the linguistic type of the word: tense, sections of discourse, plural or solitary, sexual orientation and others;

c. A talked corpus is incorporated together with the interpretations of the oral writings, which enables the clients to distinguish the elocution of the words;

d. A parallel corpus is the primary hotspot for interpretation purposes, in which we can discover reciprocals of the words;

e. A corpus of the dialects gives an incredible number of cases, by which the lexicon creators can accept what might as well be called the passage word in their local dialect;

f. Encyclopedic data; all corpora composes can supply with the cases in an a lot of numbers through which one can arrive at the last conclusion;

g. Etymology of the passage words; just this foundation isn't appropriate if the corpus isn't uniquely composed including etymological lexicons and related writings.

h. A corpus is specific to discover morphemic and arrange units of the words. The corpora give more solid information with correct measurements and it spares the time and vitality to make scans for illustrations.

Now, concerning the nature of bilingual dictionaries, Pinchuck stated that the translator should bear in mind that:

(a) a dictionary, and therefore also a bilingual dictionary, is always out of date: *that means the source which is always up to date is needed, that is called a corpus of a language;*

(b) many of the recorded expressions are no longer in common use: *but, corpora are the collections of the naturally occurring texts which are generally recognized among the directly chosen individuals in their native lands;*

(c) expressions referred to as colloquial or non-standard may have risen into more formal use: *whereas, the corpus gives the data and statistics and it doesn't judge itself, it allows the users to do so;*

(d) most commonly, new expressions have come into use but are not yet recorded. *However, corpora are available to be updated daily.* As can be seen, we have indicated that a language corpus is a powerful tool that can be a very good solution for most of the mentioned problems in the sphere of dictionary-making as well.

As we have stated above, the role of corpus analysis is being marked with very high appreciations by the scholars

in different aspects of linguistics. In fact, the lexicographers are those who can indicate the real value of corpora, because they use the data supplied by the corpora in every day of their work as the source of examples and reliable evidence. That's why, the application of a language corpus in dictionary making should be taken into account consciously.

In order to show the great difference in value and efficiency of corpus-based dictionaries comparing to the traditional ones we have taken a single word that is '*fabric*'. Here, we have taken several analyses on this word in several dictionaries which we own now, at the same time, we have tried to compare the latter with the former. So, we are presenting this process of comparison accordingly:

The first dictionary we have chosen was *the Cambridge Advanced Learner's Dictionary, Cambridge University Press 2008*. According to this dictionary we could find the following data concerning the word '*fabric*':

fabric /'fæb.rɪk/ noun **CLOTH**

1. [C or U] (a type of) cloth or woven material

dress fabric

seats upholstered in hard-wearing fabric

cotton fabrics

fabric /'fæb.rɪk/ noun **STRUCTURE**

2. **the fabric of sth**

the structure or parts especially of a social unit or a building

the fabric of society

Unhappiness was woven into the natural fabric of people's lives.

We must invest in the fabric of our hospitals and start rebuilding them.

Linen is a coarse-grained fabric.

The fibres are woven into fabric.

This fabric is similar to wool, only cheaper.

The designs are printed onto the fabric by hand.

She designs her own fabrics using woodblocks and stencils to create patterns on the material.

However, in our opinion, the dictionary should provide the user with more information about the word forms, parts of speech in one search and, of course, more examples ought to be provided.

The next dictionary was *Oxford Dictionary of English, Revised Edition, Oxford University Press 2005*. This dictionary also supplied the readers with comparable data about the word '*fabric*':

fabric [mass noun]

1) cloth produced by weaving or knitting textile fibres heavy cream fabric | [count. noun] stretch fabrics

2) the walls, floor, and roof of a building

■ the body of a car or aircraft

■ the basic structure of a society, culture, activity, etc
the multicultural fabric of Canadian society

• Origin: late 15th cent.: from French *fabrique*, from Latin *fabrica* 'something skilfully produced', from *faber* 'worker in metal, stone, etc.' The word originally denoted a building, later a machine, the general sense being 'something made', hence sense 1 (mid 18th cent., originally denoting any manufactured material). Sense 2 dates from the mid 17th cent

This dictionary explains the meaning of the word very well, however, lack of examples cannot give a general view of understanding it. It means that the users are not provided with enough examples as if only theory but not practice. No synonyms or antonyms are mentioned that means the learners cannot acquire sufficiently.

The third dictionary being an example for traditional dictionaries was *Collins Cobuild Advanced Learner's English Dictionary, New Digital Edition 2008*.

fabric [fæbrɪk] fabrics

1) N-MASS: Fabric is cloth or other material produced by weaving together cotton, nylon, wool, silk, or other threads.

Fabrics are used for making things such as clothes, curtains, and sheets.

...small squares of red cotton fabric... Whatever your colour scheme, there's a fabric to match.

2) N-SING: *with supp. usu of the N of n*. The fabric of a society or system is its basic structure, with all the customs and beliefs that make it work successfully.

The fabric of society has been deeply damaged by the previous regime... Years of civil war have wrecked the country's infrastructure and destroyed its social fabric.

3) *usu the N of n*. The fabric of a building is its walls, roof, and the materials with which it is built.

Condensation will eventually cause the fabric of the building to rot away.

We found this dictionary well-explained with few examples and there is no word about the origin of the word and nothing is told concerning the word formation and synonyms as well as antonyms.

Now, let's make some searches in several corpora including BNC, COCA ([Corpus of Contemporary American English](http://www.wordandphrase.info/frequencyList.asp)) and GOOGLE BOOKS: BRITISH ENGLISH. The first, we need to identify the meaning of the word. Here, we made a search in COCA, it provided us with lots of important data about 'fabric'.

From COCA we have taken the definition given in this link, <http://www.wordandphrase.info/frequencyList.asp>: that is shown in the Picture 1.

Picture1

DEFINITIONS (WORDNET)

1. artifact made by weaving or felting or knitting or crocheting natural or synthetic fibers 2. the underlying structure

COLLOCATES (click to see with FABRIC)

adj: social, very soft, synthetic, thin, woven, space-time, natural, urban, luxurious **noun**: piece, part, color, society, cotton, design, silk, store, pattern, thread **verb**: weave, cover, cut, create, choose, tear, buy, pull, stretch, drape

(BAD ENTRY?)

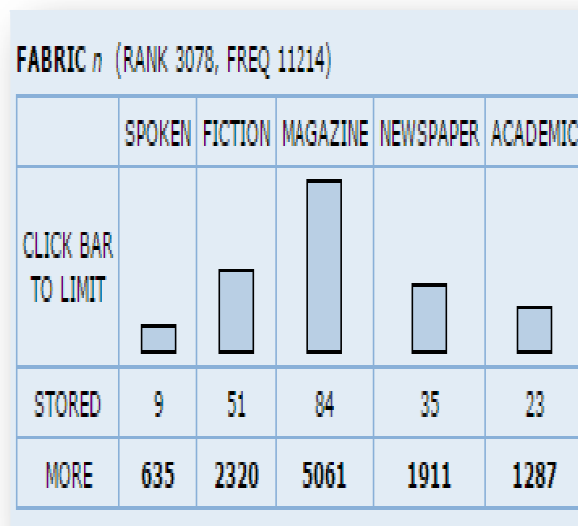
Moreover, it provided us with the information about the word formation of the word 'fabric' in one table as it is shown in the Picture 2. Here, the signs have their specific meanings, that is, N-noun, V-verb, ADJ-adjective. A glance at the given table supplies the striking reveals that N-noun form of the word is the most common among the others, indicating the total amount of 11,214 examples. Another very important feature of this table is that it allowed us to find examples according to the genre of usage. One can make an assumption that the word 'fabric' is mostly used in magazines comparing to the rest of the genres. That can also be seen in the picture 3 in the form of bar graph. There, the column of the magazine is much higher than that of the rest.

Picture 2

CLICK ON A WORD FOR CONTEXT AND SYNONYMS (CAN ALSO LIMIT TO GENRE)

	RANK #	POS	WORD	TOTAL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
1	3078	N	FABRIC	11214	635	2320	5061	1911	1287
2	10618	V	FABRICATE	1511	199	126	403	238	545
3	11783	N	FABRICATION	1296	125	114	262	152	643
4	28331	J	FABRICATED	187	15	22	47	32	71
5	32366	N	FABRICATOR	146	10	11	44	34	47
6	51714	J	FABRIC-COVERED	32	1	8	15	7	1

Picture 3



The next picture (Picture 4) taken from the corpus has provided us with the data collections with the list of collocations with the word 'fabric'. From here, the learners and lexicographers can learn a lot about the word and they would use it in their practices.

Picture 4

CONCORDANCE LINES

CLICK WORD TO

SEARCH HAS COLLOCATE

QUERY THAT WORD

2

GENRE		WORD	WORD	WORD
1	WAG	floral print on the side. (Keep a room's style fabric and accessories simple when such a bold print sits center stage.		
2	WAG	with a sunny smile of her own. Photograph for fabric and accessory sources turn to page 115. (Sticker LVM's		
3	WAG	together an entire house from the back of a seat. Fabric and carpets for \$15 to \$60 per yard installed, and		
4	FTC	as a high-school guidance counselor, fabrics ready-to-wear fabric and customly never bought -- let alone worried about -- cloth		
5	NEWS	Ugly, / (You sent out an array of opportunities for fabric and colors that only someone with his talent can mix. (How		
6	FTC	the highly padded hangers, close enough to feel the difference fabric and recent pieces the dresses to the needs of high heels.		
7	NEWS	higher-performance engines. = Interior. Better-looking fabric and accessories look better buckle make the piece nicer. A		
8	ALAD	which they are set, substituted plain cotton for decorative fabric and measured out a shade (Thanks for the lady's long, feminine,		
9	WAG	a living room. We used a simple treatment of white fabric and red velvet lace on the plain, three-style window. For a		

The following table below was taken from COCA which provides the synonymic row of the word 'fabric'. In that table, the synonyms are presented with the dominant word according to the meaning and the frequencies of the words. The word fabric can be seen in three categories of meaning: *brickwork*, *cloth* and *structure*. (Table 3)

Table 3. SYNONYMS

brickwork	
951	structure
3078	fabric
15431	masonry
16300	roofing
22905	superstructure
29226	tiling
29658	stonework
33071	cladding
32246	facing
34149	brickwork
Cloth	
997	stuff

3078	fabric
3566	cloth
5688	textile
15558	drapery
Structure	
951	structure
1843	frame
2270	foundation
2987	framework
3019	composition
3078	fabric
3388	makeup
4504	constitution
6725	basics

Now, let's learn about evolution of the word 'fabric' due to the numbers of use. Here, we have decided to utilize the Corpus of GOOGLE BOOKS: BRITISH ENGLISH. When the search was done, we have clicked on the button *select*, and identified the possibility of categorization of the word usage according to the years. The results are given as following in table 4.

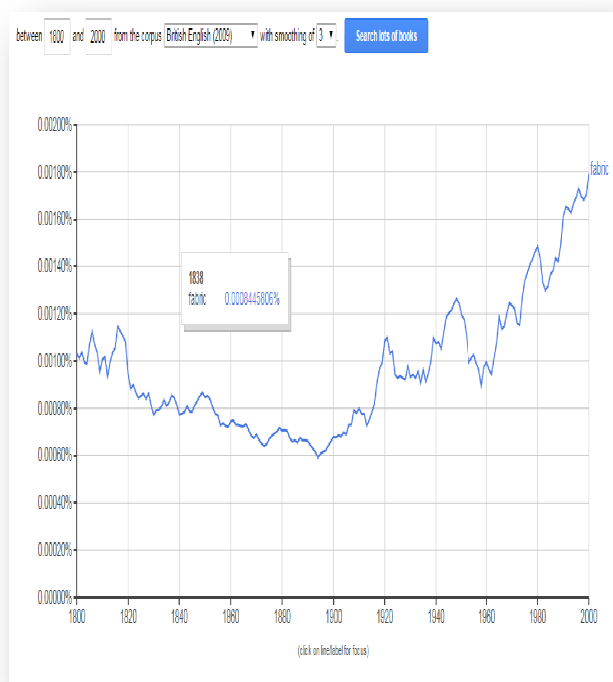
Table 4. Word: fabric

Year	Number
1810	8174
1820	11088
1830	11086
1840	13570
1850	15539
1860	13466
1870	15153
1880	17816
1890	18725
1900	22523
1910	10947
1920	19237
1930	16647
1940	15458
1950	22266
1960	30225
1970	35556
1980	43541
1990	71830
2000	183062

Total:	595909
---------------	---------------

From this information we could get the results of comparison, that is, at the beginning of the XIX century the word of '*fabric*' wasn't used very much and upto that time no examples given in the corpus. That means this very period is the origin of the word and the development of the usage can be seen in narrow scale easily: at the beginning there were 8174 examples of the very word use, and in 2000 it became 183062. This great difference is described in picture 5. This line graph also was provided by the corpus. A glance at this graph provides information about that the application of the word '*fabric*' has developed through years.

Picture 5



Here, we have covered the information about the efficiency of corpus application in lexicography. As it is stated above, the present day dictionaries are not perfect, that is, each of them should be improved and updated daily. However, it is very hard job to do if the lexicographer doesn't use corpus, because corpus is able to provide the definition, collocations, word formation, evolution and, of course, unlimited number of examples. So, we consider the application of the corpora in compiling dictionaries very vital.

REFERENCES

- [1] Nikola DobrićMisic, Lopovic (Eds.) Language, literature and identity, pp. 359-363, 2009.
- [2] Meyer, C.F. (2002). English Corpus Linguistics. Cambridge: Cambridge University Press.
- [3] Lindquist, H. (2009). Corpus Linguistics and the Description of English. Edinburgh: Edinburgh University Press.
- [4] Cambridge Advanced Learner's Dictionary © Cambridge University Press 2008 Produced for Cambridge University Press by Armada.
- [5] Collins Cobuild Advanced Learner's English Dictionary. New Digital Edition 2008 © HarperCollins Publishers 2008
- [6] Oxford Dictionary of English, Revised Edition. © Oxford University Press 2005.
- [7] Available [Online] at: http://googlebooks.byu.edu/help/intro_e.asp?w=&h=
- [8] Available [Online] at: <http://corpus.byu.edu/coca/>
- [9] Available [Online] at: <http://corpus.byu.edu/bnc/>