ISSN 2181-922X

UZBEKISTAN

LANGUAGE & CULTURE

O‘ZBEKISTON

TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2023 Vol. 1 (6)

www.compling.tsuull.uz

MUNDARIJA

ISSN 2181-922X 6-17

CORPUS FOR WHAT

Eşref Adalı[1]

Abstract.

A corpus can be called a collection of words and phrases that are created to process a language in general. A competent corpus must be large and have all the features of the language. The corpus can be divided into two as ***"Balanced"*** and ***"Unbalanced".*** The unbalanced corpus is the large corpus. It contains many texts and the source of these texts is not important. A balanced corpus is created by taking equal lengths of text from all fields in that language. We can divide the corpus into two classes according to their purpose: In terms of grammar and Natural Language Processing. The corpus that should be prepared to evaluate the developments in a language should be a balanced corpus. The texts to be included in the balanced corpus should represent that language in the best way. This type of corpus is mainly used by linguists. Corpus size must be as large as possible for NLP studies. It is appropriate to select the texts to be included in the compilation from different fields, but it is not so important to balance them.

Keywords: *Corpus, Parallel Corpus, Balanced Corpus, Unbalanced Corpus, Language Modelling, Zift Law.*

1. Introduction

A corpus can be called a collection of words and phrases that are created to process a language in general. The corpus is defined differently by different scholars:

- *"Corporation is a collection of linguistic information, which may consist of written text or recorded conversations, used to make assumptions about a language or to begin a linguistic description of the language."* [Crystal, 1991].
- *"A naturally occurring repertoire of texts chosen to represent the characteristic feature of a country and the diversity of its language."* [Sinclair, 1991].

Corpus can also be defined as a special database consisting of texts that can be used in the field of "Natural Language Processing"

[1] *Eşref Adalı* - doctor of technical sciences, professor. Computer Eng. & Informatics Faculty İstanbul Technical University Istanbul, Türkiye.
E-mail: adali@itu.edu.tr
ORCID: 0000-0002-1561-8255

and can provide fast and accurate operations on words.

A competent corpus must be large (containing a large number of words) and have all the features of the language. Corpus can be created in two ways:

- **Quality Corpus:** It is a corpus that shows the features of the language and includes examples from the texts in the language.
- **Large Corpus:** Contains more text for use in Natural Language Processing.

In addition, the corpus can be divided into two as ***"Balanced"*** and ***"Unbalanced"***. The unbalanced corpus is the large corpus. It contains many texts and the source of these texts is not important. A balanced corpus is created by taking equal lengths of text from all fields in that language. However, finding equal-sized text from all fields is a difficult process. The unbalanced corpus can be used in different fields because it contains more text. If the aim is to do letter analysis, a small corpus is sufficient [Dalkılıç, 2001]; however, if lexical analysis is to be done, a very large corpus is required. Also, for some unusual words, unbalanced corpus is more useful.

The created corpus may be an example of the current written language, or it may consist of old books or documents or speeches representing the spoken language [Church, Mercer, 1993]. In a language, the number of words used in oral expression is less than the number of words used in written expression, in addition, the word structure in oral expression may vary according to written expressions due to dialect differences or other reasons [Jurafsky, 2000].

Unlike the written corpus, the oral corpus usually does not contain punctuation marks, but it may also contain words that are uncertain whether to be processed as words. Words can be left unfinished, and verbal (such as hı, hım) and nonverbal (silence) pause expressions can be found that are not in the written corpus. In addition, each of these expressions has its own meaning. These meanings should also be investigated and it should be determined whether these words are specific to that language and can be included in the corpus.

During the creation of the corpus, it should be determined how to evaluate the words that derive from the same root, such as compound words and plural words, but that can also contain different meanings. The fact that compound words or plural words can be evaluated as separate words in the corpus will affect the number of words that make up the corpus, and will bring about various changes in the analysis algorithms or the development of algorithms that take into account all possibilities.

Corpora (or corpuses) are prepared for different purposes. In this article, it will be explained what kind of corpus is appropriate for which purpose.

- To evaluate the developments in the language,
- Correcting spelling mistakes,
- For grammatical analysis of language
- Correcting grammatical errors
- Translating speech into text
- Voice the text
- The meanings of sentences can be deduced,
- Assistance with encryption and decryption,
- For machine translation.

We can divide the corpus into two classes according to their purpose: In terms of grammar and Natural Language Processing

1- To evaluate the devolepments in the language

Languages change over time. The corpus makes an important contribution to the monitoring of the developments in the language in certain time periods. It is known that there are important changes in Turkish as a result of the innovation studies that started with the language revolution. Similar developments will be valid for other Turkic (Uzbek, Kazakh, Turkmen, Kyrgyz, Tatar, Yuvash, etc).

The corpus that should be prepared to evaluate the developments in a language should be a balanced corpus. The texts to be included in the balanced corpus should represent that language in the best way. This type of corpus is mainly used by linguists. Linguists want to use the corpus for the following studies:

- To reveal the vocabulary of the language and accordingly to prepare a dictionary.
- To determine the time a word was born and to examine its frequency of use over time.
- Watching a word get forgotten over time.
- The ratio of native and foreign words in the language and the changes in the ratios, to evaluate the effect of foreign languages.
- Preparation of the frequency dictionary.
- Identifying words used by an author.

A balanced corpus is created by taking samples of a certain length from the texts published in the specified time period in the selected language. The aim is to prepare the corpus that will represent the language in the most capable way. In the first stage, the topics, authors and sample text sizes are determined. The use of the language in different areas is exemplified by selecting the texts from different areas. For example, daily news, educational books, current information sources, texts from fictional articles are compiled. Sub-headings are created under each topic. Before starting all these studies, the weight of each subject and sub-topic in language representation is calculated. The weight of a subject in the representation of language depends on the importance that society attaches to it. In a sense, the frequency of reading the texts published on this subject indicates the weight coefficient of the subject. The size and number of samples to be selected are related to the size of the corpus that is aimed to be prepared. For balanced corpuscles, the size of the corpus is expected to be 2.500.000 words and the sample text size to be 5000 words. Based on these numbers, the number of sample texts is calculated. In Table-I, the topics that we can suggest for the generic corpus, the weight of the topics and the number of texts are given. The total number of texts in our proposed corpus is 500. Therefore:

Corpus size = 2.500.000 = 500 x 5000 words

The procedures to be performed for each text included in the compilation are as follows:

• 500 works representing the language must be determined within the specified time period.

• A portion of 5000 words should be chosen randomly from the determined works.

• Imprint should be prepared for each selected text. The name of the work, its author and time information should be included in the imprint.

The size of the corpus can be enlarged in accordance with the basic principles outlined above. For example, if the number of texts is increased to 5000, the size of the corpus becomes 25 million words. Such a corpus can be considered quite competent for the study of the Turkish language. Naturally, it is difficult to prepare a balanced collection of this size. Since the publications published in the past are not in the digital environment, the sections selected in the publications should be scanned with optical readers and transferred to the information environment. At the end of the scanning process, it cannot be said that the texts are converted into characters without errors. Depending on the printing and paper quality, it can be said that the success of the scanning process will be between 80-90%. Therefore, mistakes must be corrected by humans or by NLP methods. Assuming that a book page contains an average of 350 words, it can be said that a text consisting of 5000 words is approximately 14 pages. For a collection of 5000 texts, 85,000 pages of text must be scanned. As can be understood from this simple calculation, the preparation of the balanced corpus requires labor-intensive work.

For each text included in the balanced corpus, an imprint containing the following information should be prepared:

- Name of the writer
- Name of the work or article
- Type of work or article
- Release date
- Printing place
- Target audience

In addition to all this, it may be necessary to respect copyrights for selected texts.

2- NLP purpose corpus

The corpuscles to be used in NLP studies do not need to be balanced corpus. Therefore, the unbalanced corpus will be introduced.

Unbalanced Corpus

It is necessary for such studies to be as large as possible, which will be required for NLP studies. It is appropriate to select the texts to be included in the compilation from different fields, but it is not so important to balance them. Those working in the field of NLP want to use the corpus for the studies listed below.

- Revealing the sound features of the language.
- Revealing the morphological features of the language.
- Determination of the class of words.
- Determination of spelling principles, accordingly finding spelling mistakes.
- Determination of word meanings.
- Revealing the syntactic features of the language.
- Modeling the language.

Table-I: Topics, Number of Text and Influence for Generic Balanced Corpus

Type of text	Number of text	influence factor %	Total inf. factor %
Articles in the press			**18**
News	***47***	***9,4***	
• Political	10	2	
• Sports	10	2	
• Social	12	24,4	
• Daily	5	1	
• Financial / Economics	5	1	
• Cultural	5	1	
Columns in papers	***22***	***4,4***	
• Columns	12	2,4	
• Daily comments	5	1	
• Editorial	5	1	
Comments	***21***	***4,2***	
• Theater	5	1	
• Book	6	1,2	
• Music	5	1	
• Art	5	1	
Educational publications			**37**
Scientific publications	***150***	***30***	
• Science	25	5	
• Mathematics	10	2	
• Technical and engineering	20	4	
• Social sciences	45	9	
• Medicine	10	2	
• Political sciences, law	20	4	
• Education	20	4	
Religious publications	***10***	***4***	
Publications on hobbies and occupation	***15***	***3***	
Current information publications	**40**		**8**
• Information publications	15	3	
• Parliament minutes	10	2	
• Corporate announcements	10	2	

• University theses	5	1	
Fiction	**140**		**28**
• Novel (general)	40	8	
• Stories	25	5	
• Crime	25	5	
• Science fiction	25	5	
• Adventure	25	5	
Real articles	**30**		**6**
• Souvenir	10	2	
• Travel writings	10	2	
• Private letter	5	1	
• Trial	5	1	
Humor	**15**	**3**	**3**
Total	**500**		

In the preparation of the corpus, which will form the infrastructure for the above-mentioned purposes, the publications on the web and the books and documents prepared in the informatics environment can be used. With the texts to be collected from different sources, corpuses that can exceed 100,000,000 words can be created. The corpus required for NLP research,

- It may consist of texts without imprint information.
- Words in the corpus may be tagged with their attributes.
- Roots of words, their classes, the links between them and their meanings may be specified.

Depending on the aim of the study, one of the above corpora can be selected. It can be said that human effort will not be enough to prepare the third most comprehensive and talented corpus. It is extremely difficult to label 100 million words individually as roots, classes, links and meanings. But such qualified corpus may be required for NLP studies. It has been seen that the smaller ones were prepared by people in the past years. Such a study was carried out by Kucera and Francis at Brown University in 1967 [Francis, Kucera, 1979]. The characteristics of this compilation, which is referred to as the Brown Corpus, are as follows:

- Selected language: American English
- Number of texts: 500
- Text size: 2000 words
- Word count: About 1,000,000

The content of Brown's corpus when it was first prepared is shown in Table-II. In this corpus prepared entirely by humans, words are labeled according to their qualities. Today, rule-based, probability-based and learning-based methods are used for labeling words. Thus, those working in the field of NLP develop a useful tool for themselves by using the opportunities provided by NLP.

Table-II: Topics, Number of Text and Influence of Brown Corpus

Type of text	**Number of text**	**influence factor %**
News	44	8,8
• Political, Sports, Social, Sports news, Economics, Cutural		
Columns in papers	27	5,4
• Columns, Daily comments, Editorial		
Comments	17	3,4
• Theater, Books, Music, Dans		
Educational publications	80	16
• Natural Science, Medicine, Mathematics, Social sciences, Political sciences, Humanity, Technical		
Religious publications	17	3,4
• Books, Magazine, Treatise		
Publications on hobbies and occupation	36	7,2
• Books, Magazine		
Fiction	29	5,6
• Novel (general), Stories		
Memory	75	15
• Books, Magazine		
Various	30	6
• Official documents, Corporate reports, Industrial reports, University Catalogues, Industrial documents		

Unbalanced computer translation is used in the language model of the corpus, removing word and sentence ambiguities and studies. Unbalanced computer translation is used in the language model of the corpus, removing word and sentence ambiguities and studies. Therefore, we will also address these issues.

Parallel Corpus

It is the corpus type used for interlingual translation. A parallel corpus is prepared for two or more languages. A text in the first of the parallel corpus prepared for two languages is translated into the second corpus. Studies done in parallel corpus are:

Alignment of texts: The correspondence of the text in the first corpus is marked in the second corpus.

Alignment of sentences: The correspondence of the sentence in the first corpus is marked in the second corpus

Alignment of words: The correspondence of the word in the first corpus is marked in the second corpus

Since it will take a lot of time to do the alignment studies manually, methods are being developed for these studies. These methods make use of artificial intelligence methods.

Language Model

The probability distribution of the order of words in a language is defined as the statistical language model (or language model for short) of that language. Various methods have been developed to extract the language model of a language. The names of some of the methods used for the language model are given below:

Zift Law: Not all words in a language are used with the same frequency, some are used very often, while others are used very infrequently. A study on this subject has been done by George Kingsley Zipf and is referred to as Zipf's law. Zipf's law describes the relationship between the order of use and the number of uses (frequency) of words in a language. This definition does not give an equality but a certain approximation. According to this law, n represents the frequency of a word and r is the sequence number of the word [Corral, Boleda, 2017]. Accordingly, it can be written as

$$n(r) \propto \frac{1}{r^{\alpha}}$$

Here a is α constant value and $\propto$ is the sign of proportionality. Zipf's law can be interpreted as follows: Commonly used words in a language are usually functional words. For example, and, a, this, da, de, are functional words for. Some words are used infrequently. The set between frequently and rarely used words is the set of main words. The result of the Zift law is shown in Fig.-1

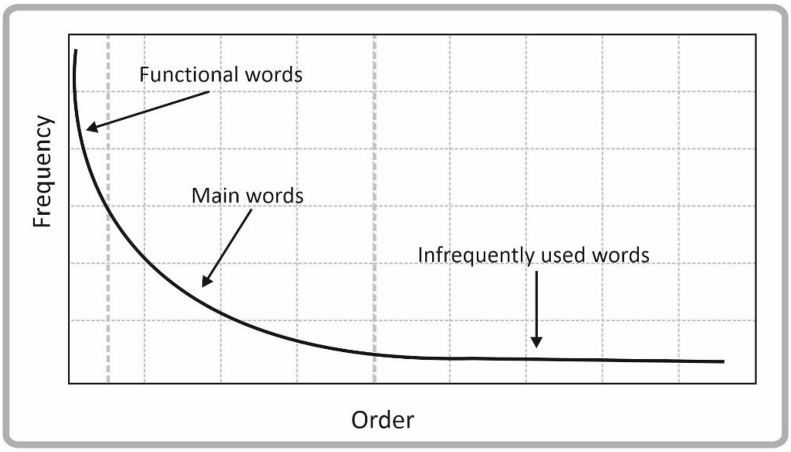

Fig. 1. The result of the Zift law

N-Gram Model, Markov Chain: This method uses previous letters or words to find later letters and words. This method is commonly referred to as the n-gram method. If the sentence is correct (written according to grammatical rules), the next word can be predicted by probability methods. This is easier in languages with strict syntax. It is a little more difficult in flexible languages like Turkic languages. A person's ability to predict the next word is the result of knowing the principles and features of his/her language. Here are the principles and features to know for this example:

Subject knowledge or world knowledge: Having knowledge about the subject of the text. To know that the text is related to language. Therefore, it can predict the next word.

Phonology: Knowing the phonetics of the language guides them to find spelling mistakes. Especially in languages with vowel and consonant harmony such as Turkish, these features make an important contribution to revealing and correcting spelling mistakes. There is no possibility to apply this feature to other languages.

Vocabulary knowledge: Must know lexicon related to linguistics. Among these lexicon, he/she can choose the appropriate one. In a sense, it limits the number of lexicon he can choose.

Syntax knowledge: Using syntax knowledge, they can decide on the order of affixes and words in languages such as Turkish.

One of the methods developed for the computer to make the prediction that human can make is the N-gram method and this method;

- Finding spelling mistakes
- Finding missing words
- Labeling words
- In computerized translation

is used. The basis of the method is that the missing word can be predicted in accordance with probability methods based on the words preceding it. The missing word is referred to as a one-gram based on the word preceding it, a two-gram based on two words, and a three-gram based on three words.

The Markov hypothesis states that the probability of the missing word depends only on the words that precede it. According to this assumption n. The probability of the word depending on the words before it is written as:

$$P(S_{ö}|S_a \cdots S_n\text{-}1) = P(S_n|S_{n-k}\, S_{n-k+1} \cdots S_{n-1})$$

According to the Markov hypothesis, we can write the n-gram probabilities as:

one-gram: $P(S_1\, S_2 \cdots S_n) = P(S_1)\, P(S_2) \cdots P(S_n)$
two-gram: $P(S_1\, S_2 \cdots S_n) = P(S_1)\, P(S_2|S_1) \cdots P(S_n|S_{n-1})$
three-gram: $P(S_1\, S_2 \cdots S_n) = P(S_1)\, P(S_2|S_1) \cdots P(S_n|S_{n-2}\, S_{n-1})$

References

D. Crystal, A Dictionary of Linguistics and Phonetics, Blackwell, 3rd Edition. (1991)

J. Sinclair, Corpus Concordance, Collocation. OUP. (1991)

G. Dalkılıç, Some Statistical Properties of Contemporary Printed Turkish and A Text Compression Application. MSc Thesis. International Computing Institute, Ege University. (2001).

K. Church, & R. Mercer, Introduction to the Special Issue on Computational Linguistics Using Large Corpora. Computational Linguistics, 19:1, pp. 1-24. (1993).

D. Jurafsky, & J. H. Martin, J.H. Speech and Language Processing, Prentice Hall, pp. 193-199. (2000).

W. N. Francis & H. Kucera, Brown Corpus Manual, Brown Univ. 1964,1971, 1979

Á. Corral, G. Boleda, R. Ferrer-i-Cancho, Zipf's Law for Word

Frequencies: Word, Forms versus Lemmas in Long Texts, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4497678/ [24.12.2017 18:32:57]