STATISTICAL ANALYSIS OF COLLOCATIONS OF THE CONCEPT JOY IN R. IVANYCHUK'S TEXT CORPUS

Nataliia Lototska
Post-graduate student, National University "Lviv Polytechnic",
e-mail: nata07lototska@gmail.com, orcid.org/0000-00016692-196X, Ukraine

Abstract. The paper includes a review of scientific works on the importance of corpus and quantitative methods, the problem of connectivity and the ways of collocation study. The article deals with the study of collocations of the emotion JOY in writer's Text Corpus by the means of statistical methods in modern linguistics. From the point of view of language system described collocations are presented in various structural-semantic forms in author's idiolect. Meanwhile statistical research represents a list of collocations organized according to absolute and relative frequency and association measures such as T-score and MI-score.

Keywords: Text Corpus, collocation, node, collocate, statistical analysis, frequency, association measure.

DOI: http://dx.doi.org/10.23856/3709

Introduction

The artistic language reflects not only the linguistic competence of the author, and the advantages of using one or another language constructs and words over, but also features of the national language (*Kulchystkyi, 2017*). Quantitative analysis is used to study author's style to avoid methodological mistakes frequently caused by researcher's subjectivity in giving examples for a suggested hypothesis.

Corpus and statistical approach in linguistic research

Development of Text Corpus leads to increased efficiency in linguistic processing of large text databases. Text corpus provides great opportunities for conducting various linguistic studies of the language system. A corpus is "a collection of pieces of language text in electronic form" (*Sinclair, 2004: 19*), meanwhile, text is "natural language used for communication, whether it is realized in speech or in writing" (*Biber & Conrad, 2009: 5*). Corpus linguistic research offers strong support for the idea that language variation is systematic and can be described using empirical, quantitative methods. Text corpus is used to perform statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules. Text corpus is electronically processed in text analytical tools and possesses useful statistical information such as number of word types, frequency, co-occurrences (*Biber & Conrad, 2009*).

Statistical methods are important and reliable tool for linguistic data analysis in modern linguistics. In addition, quantitative methods ensure reliability of results, allow to reveal language units and text structure properties, the research that would be impossible without statistical studies. The fact that language itself is a complex system subordinated to the laws of statistics proves the necessity of using statistical methods in linguistics (*Perebyinis, 1967*).

Collocation as lingual / language system unit

The object of our study is a collocation in R. Ivanychuk's Text Corpus. There are different approaches to definition of the term 'collocation'. Sometimes the 'collocation' is used as a synonym of a word combination, sometimes as a special type of a set phrase. In corpus linguistics the 'collocation' is the word combination used in the text together more often, than used at random probability separately, in other words collocations are understood as statistically determined set phrases. S. Evert suggests the following definition: "A collocation is a word combination whose semantic and syntactic properties can't be fully predicted on the basis of information about its constituents and which therefore should be added to the dictionary (lexicon)" (*Evert, 2004: 17*). The text corpus and tools of corpus linguistics make possible to identify and expand the lexical fund of set phrases of various types and peculiarities of their use (*Zakharov, 2015*).

It is known that "the language system is probabilistic, and frequency in a text is an illustration of grammatical probability" (*Halliday, 1991: 31*). This suggests that words in speech are subordinated to grammatical rules of language and aren't used arbitrarily in a language flow.

One of the main approaches of working with corpus data is to study collocations is concordance — Text Corpus lines representing the word in context. Concordance lines are the source of information about patterns of usage of word (node) and the connection between other words (collocate).

Statistics to study collocations in Text Corpus

In computational linguistics the term 'collocation' is defined as 'statistically stable word combination" (*Khokhlova, 2010: 8*). The most basic corpus-based statistics are the absolute frequency and the relative frequency of some phenomenon. *The absolute frequency* (co-occurrences) is a number of times that a value appears, the sum of the absolute frequencies is equal to the total number of word types in Text Corpus. At the same time the *relative frequency* is an estimate of the probability of a given phenomenon in the language.

In addition, collocations are studied by means of mathematical criteria — statistical association measures, which are based on probability theory and mathematical statistics. Association measures are mathematical formulas determining the strength of association between two or more words based on their occurrences and co-occurrences in a text corpus. It is known that *T-score* extracts most frequent collocations. On the contrary, the *MI-score* allows to reveal low-frequency multiword terms and proper names. These measures play an important role in the automatic extraction of collocations.

Lexical association measures are applied to a key word (node) occurrence and context statistics extracted from the corpus for all collocation candidates and result in their association scores. On the top of the list are word combinations that are assumed to have the greatest association with each other and, consequently, be the most probable collocation candidates. The frequency of joint occurrence of a key word (node) and its collocate is taken into consideration (*Zakharov, 2015*).

Statistical methods allow to obtain reliable statistics data of lexical unit compatibility based on Text Corpus, to study lexical units in context, to obtain data on frequency of words, lemmas, grammatical categories, co-occurrences of lexical units, compatibility peculiarities. In addition, search results can be ranked by different parameters and we are able to set

threshold values making possible to obtain meaningful information (*Khokhlova*, *2010:* 66). The co-occurrence is associated with the frequency of individual components of the collocation.

Statistical research of collocations with the concept JOY in R. Ivanychuk's Text Corpus

Author's vocabulary research allows to describe the lexical arsenal of writer's idiolect and will make possible to identify his texts among others. Words marking emotion are one of the key factors in the comprehension of author's language and his personality.

Valuable information about idiolect specificity is represented by word frequency analysis. Collocation study provides important information about author's style peculiarities but collocation research in fiction is insufficiently studied in Ukrainian linguistics.

Corpus for the study contains texts by R. Ivanychuk, a Ukrainian writer (1929-2016); 16 historical novels and 1 historical trilogy (1962-2016) with 1,295 million words analyzed («Край битого шляху», «Мальви», «Черлене вино», «Манускрипт з вулиці Руської», «Вода з каменю», «Четвертий вимір», «Шрами на скалі», «Журавлиний крик», «Бо війна війною», «Орда, Євангліє від Томи», «Вогненні стовпи», «Саксаул у пісках», «Через перевал», «Хресна проща», «Голоси з-над вод Генісарета», «Я ще не писав про Донбас»).

In this study word combinations with the concept JOY are described and extracted by means of the Ukrainian corpora 'GRAK' and Collocation tool of the NoSketch Engine system, association measures such as T-score and MI are used to study collocations.

In R. Ivanychuk's idiolect the collocations were described and analyzed by means of absolute and relative frequency, association measures T-score and MI-score. In our research the following frequencies are taken into consideration: ≥ 2 frequency for MI-score, and ≥12 frequency for for T-score. The absolute frequency of lemma JOY is 278 word types (node) and the relative frequency is $2{,}15*10^{-4}$ in R. Ivanychuk's Text Corpus.

The most frequent collocations correspond to high frequency constructs according to the T-score: *і радість / радість і, від радості, велика радість, з радості / з радістю, радість від, для радості, сльози радості*. These word combinations are used by the writer subconsciously and serve as a basis for identification of author's texts.

Collocations extracted by MI-score are the less frequent combinations, in their turn, they are individual set phrases illustrating the author's idiolect and can be served as indicator of writer's text attribution: *затеплилася радість, підленька радість, скритна радість, незмірна радість, притлумлювати радість* etc.

The study of collocations with the concept JOY in R. Ivanychuk Text Corpus

The research shows that the concept JOY is presented in various structural-semantic forms in R. Ivanychuk's Text Corpus. High frequency collocations are grammatical constructions without any special semantic coloring, but they can serve as formal indicators in the author's text study. For example, constructions with a preposition — Prep + JOY / JOY + Prep: *від радості, з радості, з радістю, за радість, для радості, на радість, від радості, у радості, після радості, радість перед, радість від, радість за, замість радості, од радості*; constructions with a coordinating conjunction — Conj + JOY / JOY + Conj: *і радість, радість і, радість й, й радість, радість чи, радість або*; constructions with a particle Partic + JOY / JOY + Patric: *ні радість, не радість, ні радості, то радість, більше радості, стільки радості*.

Table 1

Most probable collocates for JOY extracted by means of T-score

Collocates	Collocations	T-score	Absolute frequency	Relative frequency
від	від радості, радість від	4.539	22	$1.7*10^{-5}$
і	і радість, і радість	4.387	28	$2.2*10^{-5}$
з	з радості, з радістю, радість з	3.875	20	$1.5*10^{-5}$
й	й радість, радість й	3.468	15	$1.2*10^{-5}$
великий	велика радість	3.084	10	$7.7*10^{-6}$
сльози	сльози радості	2.223	5	$3.9*10^{-6}$
незмірний	незмірна радість	1.999	4	$3.1*10^{-6}$
творення	радість творення	1.997	4	$3.1*10^{-6}$
бурхливий	бурхлива радість	1.997	4	$3.1*10^{-6}$
тихий	тиха радість	1.992	4	$3.1*10^{-6}$
для	для радості	1.954	5	$3.9*10^{-6}$
передчуття	передчуття радості	1.729	3	$2.3*10^{-6}$
буйний	буйна радість	1.729	3	$2.3*10^{-6}$
новий	нова радість	1.727	3	$2.3*10^{-6}$
ставати	радість стала	1.708	3	$2.3*10^{-6}$
ні	ні радості	1.575	3	$2.3*10^{-6}$
хвилинний	хвилинна радість	1.413	2	$1.5*10^{-6}$
зблиснути	зблиснути радістю	1.413	2	$1.5*10^{-6}$
огорнути	радість огорнула	1.413	2	$1.5*10^{-6}$
приховати	приховати радість	1.410	2	$1.5*10^{-6}$
безмежний	безмежна радість	1.410	2	$1.5*10^{-6}$
шалений	шалена радість	1.409	2	$1.5*10^{-6}$
охопити	радість охопила	1.408	2	$1.5*10^{-6}$
пізнати	пізнати радість	1.408	2	$1.5*10^{-6}$
воля	радість волі	1.398	2	$1.5*10^{-6}$
перемога	радість перемоги	1.397	2	$1.5*10^{-6}$
почуття	почуття радості	1.390	2	$1.5*10^{-6}$
за	радість за, за радість	1.388	4	$3.1*10^{-6}$
справжній	справжня радість	1.377	2	$1.5*10^{-6}$
людський	людська радість	1.374	2	$1.5*10^{-6}$
побачити	побачити радість	1.333	2	$1.5*10^{-6}$
перед	радість перед	1.294	2	$1.5*10^{-6}$
чи	радість чи	1.084	2	$1.5*10^{-6}$
то	то радість	1.064	2	$1.5*10^{-6}$

Table 2

Most probable collocates for JOY extracted by means of MI-score

Collocates	Collocations	MI-score	Absolute frequency	Relative frequency
затеплитися	затеплилася радість	15.525	1	$7.7*10^{-7}$
підленький	підленька радість	14.464	1	$7.7*10^{-7}$
скритний	скритна радість	13.577	1	$7.7*10^{-7}$
незмірний	незмірна радість	13.048	4	$3.1*10^{-6}$
притлумлювати	притлумлювати радість	12.892	1	$7.7*10^{-7}$
розкаяння	радість розкаяння	12.474	1	$7.7*10^{-7}$
вмістилище	вмістилище радості	12.178	1	$7.7*10^{-7}$
зловтішний	зловтішна радість	12.130	1	$7.7*10^{-7}$
вділити	вділити радості	12.123	1	$7.7*10^{-7}$
хвилинний	хвилинна радість	12.067	2	$1.5*10^{-6}$

Prepositional phrases are divided into two semantic groups: collocations referring to CAUSE — *від радості, з радості, з радістю, за радість, для радості, на радість, від радості, од радості;* collocations referring to LOCUS — *у радості, замість радості, після радості, радість перед.* (15) Collocations including a coordinating conjunction are presented by combination of JOY with both positive and negative emotions words: *радість й печаль, радість і туга, радість й мука, біль і радість, радість й біль, злоба й радість, радість і плач, тривога й радість, здивування і радість, здивування та радість, радість і здивування, подив і радість, радість і задоволення, турбота і радість, радість і надія*.

Constructs Adj + JOY are used to describe the emotion by numerous attributes — *велика радість, бурхлива радість, тиха радість, незмірна радість, буйна радість, нова радість, безмежна радість, людська радість, надмірна радість, невимовна радість, справжня радість, шалена радість, хвилинна радість, безмірна радість, бентежна радість, весняна радість, всенародна радість, велика радість, дарована радість, дивна радість, дитяча радість, єдина радість, затаєна радість, злобна радість, материнська радість, радість миттєва, молодеча радість, млосна радість, небувала радість, несподівана радість, нестримна радість, нечувана радість, нинішня радість, остання радість, передчасна радість, перша радість, підленька радість, подібна радість, порывиста радість, прихована радість, повна радість, рання радість, сподівана радість, хвилева радість, хлоп'яча радість*. To indicate the intensity of attributives the writer uses such adjectives as: *нестримна, шалена, порывиста, незмірна, безмірна, невимовна, безмежна, надмірна, нечувана, млосна, буйна, бентежна*; to express a negative connotation of the node: *злобна, підленька*; to express the meaning of mystery: *скритна, потаємна, затаєна*; to highlight the time flow — хвилева, миттєва тощо. It is important to note that the collocation *підленька радість* is unique in the Ukrainian corpora 'GRAK' as well as in Ukrainian fiction literature.

In order to describe the emotion the writer uses phrases including verb, noun, preposition (or without preposition) V + JOY / JOY + V / V+ Prep + JOY that are typical of Ukrainian language: *плакати з радості, відчути радість, давати радість, зблиснути радістю, пізнати радість, приховати радість, побачити радість, радість охопила, радість огорнула, дарувати радість, загорітися радістю, затеплилась радість, зблискувати радістю, знаходити радість, приносити радість, подарувати радість, притлумлювати радість, радість навіювати, розділити радість, спалахнути радістю,*

пробивалися радість, приходить радість, радість діймала, радість затьмарювати, іскрилася радість, радість обнялась, радості не приносити, радість повнилася, радість пройняла, радість вщухла, радість засвітилася, радість не зблисла, радість увійшла, радість упала, радість щезла. The concept JOY may serve as a subject or as an object.

The concept JOY is observed in noun collocations in which the node can be used in nominative as well as in genitive and accusative cases: N + N: $JOY_{nom.}$ + N — *радість вивільнення, радість дозрівання, радість зачаття, радість знахідки, радість нового дня, радість очищення, радість подвигу, радість кохання, радість осяяння, радість творення, радість тріумфу*; N + $JOY_{gen.}$ — *сльози радості, передчуття радості, почуття радості, дрібка радості, радість волі, радість перемоги, вино радості, вмістилище радості, мить радості, нотка радості, плач радості, передчуття радости, приплив радості, проблиски радості, спалах радості, сподівання радости*; N + $JOY_{acus.}$ — *наповненість радістю*. The collocations with the meaning LOCUS are assigned among the described structural forms: *прихована радість, приховати радість, затеплилась радість, знаходити радість, притлумлювати радість, радість навіювати, спалахнути радістю, пробивалися радість, приходить радість*, *радості не приносити, радість повнилася, радість пройняла, радість увійшла, радість упала, радість щезла, вмістилище радості.*In linguistics a metaphor is a phenomenon of fiction literature making the author's style individual and original. Today metaphor is studied in the framework of cognitive linguistics. Cognitive linguistics research began with the publication of G. Lakoff and M. Johnson "Metaphors We Live By". In their view the essence of metaphor is understanding and experiencing one kind of thing in terms of another *(Lakoff & Johnson, 1980: 5*). G. Lakoff and M. Johnson come to a conclusion that a metaphor unites reason and imagination, creating an imaginative reality. Moreover, metaphors bring about the changes in the ways the world is perceived, and these conceptual changes often bring about the changes in the ways we act in the world, accepts Mac Cormac (*Mac Cormac, 1985: 149*). According to the new theory, metaphor is considered as a fundamental cognitive process, as a basic schema 'by which people conceptualize their experience and the external world' (*Gibbs, 1994: 1*), so the source of metaphoric language is in thought, in the organization of our conceptual system.The use of figurative expressions is observed in R. Ivanychuk's texts *від радості захлинатися, заплескати з радості, затруситися від радості, проясніти від радості, спалахнути від радості, шаленіти од радості ;до туску відчув радість вічного життя, з журбою радість обнялась, добро — то радість, м'ячик молодечої радости, радість повнилася сміливістю, радість затьмарювала думка, умитися сльозами радості, сподівання радости від праці, наповнювати вином радості, тиха радість діймала, примліти з радості, донецьке небо втішалося їхньою радістю*. Somatic metaphors are frequently used in writer's Text Corpus with such somatic markers as EYES, FACE, CHEST, SOUL, HEART: *очі зблиснули радістю, зблискують радістю очі, горять шаленою радістю очі, радість упала світлом на обличчя, в очах замерехтіла радість, очі загорілися радістю, радість засвітиться колишніми бісиками в темних очах, в очах іскрилася радість, радість не зблисла на його обличчі, налиті радістю і здивуванням очі*.

Conclusions and suggestions

Author's idiolect was studied from the point of view of language system (lexical and semantic structure) and by means of statistical parameters. As a result, it turned out that high

frequency collocations are Prep + JOY/ JOY + Prep and Conj + JOY / JOY + Conj; collocations Adj + JOY present numerous attributes; verbal and noun collocations V + JOY / JOY + V / V+ Prep + JOY / JOY + N / N + JOY demonstrate a big variety of collocates. Nominal and verbal constructs can possess direct and figurative meaning, meanwhile, among metaphors somatic metaphors are observed.

The research results are represented by a list of collocates (collocations) organized according to absolute and relative frequency and association measures such as T-score and MI-score. High frequency collocations according to absolute frequency possess high values of T-score, meanwhile, MI-score demonstrates high values for low frequency collocations. Consequently, T-score gives an opportunity to release the most frequent word combinations that can be used as formal indicators of writer's texts; MI-score allows to find individual even unique constructions that are typical of author's idiolect. Quantitative analysis used in idiolect study allows to avoid methodological mistakes frequently caused by researcher's subjectivity.

References

Biber, D., and Conrad, S. (2009). Register, genre, and style. Cambridge: Cambridge University Press. [in English].

Evert, S. (2004). The Statistics of Word Cooccurrences: Word Pairs and Collocations, unpublished doctoral dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. [in English].

Gibbs, R. W. (1994). The poetics of mind: Figurative thought, language, and understanding. New York: Cambridge University Press. [in English].

Halliday, M. A. K. (1991). Current ideas in systemic practice and theory. London: Pinter. [in English].

Khokhlova, M. (2010). The study of lexical and syntactic collocatibility in the Russian language using statistical methods (based on Text Corpus). Dissertation Abstracts, Sankt-Peterburg. [in Russian].

Kulchystkyi, I. M. (2017). The examination of sentence and word length in the writing of Roman Ivanychuk. Bulletin of the National University of Lviv Polytechnic, vol. 872, 139-149. [in Ukrainian].

Lakoff, G., and Johnson, M. (1980). Metaphors We Live By. Chicago and London: University of Chicago Press. [in English].

Levchenko, O., Romanyshyn, N., Dosyn, D. (2019). Method of Automated Identification of Metaphoric Meaning in Adjective+ Noun Word Combinations (Based on the Ukrainian Language). Lviv: National university "Lviv Polytechnic". Series Philology, Is. 70, 288–298. DOI: http://dx.doi.org/10.30970/vpl.2019.70.9784. [in Ukrainian].

Levytskyi, V. (2007). Quantitative methods in linguistics. Vinnystia: Nova Knyha. [in Russian].

Mac Cormac, Earl R. (1985). A cognitive theory of metaphor. Cambridge, MA/London, England: MIT Press. [in English].

Perebyinis, V. S. (1967). Statystychni paramentry styliv. Kyiv: Naukova dumka. [in Ukrainian].

Sinclair, J. (1991). Corpus, concordance, collocation: Describing English language. Oxford: Oxford University Press. [in English].

Zakharov, V.P. (2015). Word collocatibility in Text Corpus. Moscow: Computer linguistics and information technologies «Dialog», 14 (21), 667-682. [in Russian].