COGNITIVE STUDIES | ÉTUDES COGNITIVES, 26
Warsaw 2026

Article No.: 3319
DOI: 10.11649/cs.3319

Citation: Fokin, S. (2026). Statistical significance of indirect concept verbaliser: Case of lexeme *iodine* as a marker of disease and war. *Cognitive Studies | Études cognitives*, *2026*(26), Article 3319. https://doi.org/10.11649/cs.3319

Serhii Fokin

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
sergiyborysovych@ukr.net

Statistical Significance of Indirect Concept Verbalizer: Case of Lexeme *Iodine* as a Marker of Disease and War

Abstract

The relationship between increased frequencies of a concept in discourse and globally significant events may reflect its connection with speakers' emotional engagement. Emotionally relevant markers can be implicitly associated with an underlying concept, thereby concealing the experienced emotion. The verbaliser *iodine* can act as an indirect verbaliser of the concepts of DISEASE and WAR, which are commonly associated with fear. The article illustrates a statistical approach to processing data of the lexeme *iodine* / *ŭoɔ* frequency in internet text corpora in English (USA) and Ukrainian (Ukraine). To demonstrate the relevance of an indirect concept verbaliser and its connection with socio-historical context – in this case, the relationship between the increased frequency of the concept verbaliser *iodine* and globally relevant events, such as a pandemic, war, or a nuclear power plant accident – a statistical significance test is applied.
Keywords: concept; concept verbaliser; emotional marker; t-test; Pearson correlation; Spearman correlation

1 Introduction

Relevant and emotionally significant extralinguistic features are reflected in language and leave historical traces embedded in written texts. An increased frequency of a concept may indicate its relevance along with increased emotionality. To obtain reliable frequency data, it is necessary to use large-volume datasets. Thus, written corpora representing the language data of a community within a specific period can help uncover relevant subject-related lexemes and enable the statistical demonstration of the importance of an emotionally marked theme through the recurrent usage of direct and explicit, or indirect and secondary, verbalisers or other linguistic markers related to a core concept. For instance, terms such as *crisis*, *recession*, or *default* are direct markers of the notion of the bubble effect crisis and, consequently, verbalisers of the concept CRISIS, while lexemes such as *(un)employment*, *mortgage*, or *eviction* can serve as its indirect verbalisers, which may obscure or hide the main concept. While the former approach of establishing a correlation between a subject and an emotion seems evident and is widely used in different domains (for example, in the media, where keyword tags highlight the core subject, or in recommendation systems, which use similar key tags to suggest new adverts), the latter remains largely unexplored in philological literature. However, the phenomenon of indirect denomination of a concept is well-known in advertising:

> The products advertised are never really present in the advert, but are represented by a picture or a brand name, which metonymically stand for the item in question. On the other hand, the conceptualisation of the advertised item or service is usually expressed by either verbal or non-verbal instantiations of conceptual metaphors that act. (Velasco-Sacristán, 2010)

Modern linguistically oriented papers predominantly focus on the frequency parameters of predefined keyword sets and their variation across years (Baron et al., 2009). At the same time, in corpus-based conceptual studies, exploring the statistical significance of indirect lexical markers (i.e., indirect concept verbalisers) remains largely unexplored. Although the web space abounds with subject-focused studies, they

typically address particular aspects and do not necessarily examine the relationship between formal markers and cognitive levels. In the domain of statistical analysis of subject keywords and other lexically salient markers, Baron notes that "As such, it may surprise the reader to learn that there are relatively few studies of historical data that make use of the keywords approach" (Baron et al., 2009), a claim that remains largely valid today. While studies that rely on large-scale quantitative data for exploring cross-linguistic grammatical features are not uncommon (Fernández Zas, 2021), research addressing concept verbalisers in corpus data is comparatively scarce. Even well-known studies, such as the article "Affection, friendship, passion and charity: A history of four 'love lexemes' since the fifteenth century" by Tissari (2001), which draws on four English corpora, focus primarily on direct lexical realisations. Consequently, the systematic exploration of indirect verbalisers in relation to core concepts or keywords in the domain of corpus statistics remains limited.

The paper "Positive words carry less information than negative words" (Garcia et al., 2012) focuses on the correlation between word frequency and its emotional content, with the frequency of negative words being more relevant. Provided that *iodine* is rarely used in positive contexts and is primarily associated with the treatment of diseases, it is reasonable to consider its increased use as emotionally relevant.

Despite an impressive number of studies published on various aspects of lexemes, concepts, terms, and keyword frequency profiles, scholars rely on different methodological approaches, some of which are particularly narrow or extremely general. The purpose of the article is to test the hypothesis concerning the relationship between the usage of the lexeme *iodine/ŭoð*, as a potential concept verbaliser, and historically relevant events by applying a methodology for establishing the statistical significance of changes in frequency across years of an indirect concept marker, and to illustrate the technical procedures that can be extrapolated to other possible studies involving corpus-based data of indirect concept verbalisers.

Our working hypothesis is that the correlation between frequency oscillations in the usage of both lexemes is not due to chance. Conversely, the null hypothesis posits that the data obtained represent random values. We further argue that the revealed coincidences are due to the fact that the usage of the lexeme *iodine/ŭoð* marks crucial historical events, such as a pandemic and war, and acts as a verbaliser of the concepts DISEASE and WAR. Thus, our intention is to objectivise this subjective impression derived from statistical data and to demonstrate that the year-by-year usage rates are indeed related to significant historical events, such as the accident at the Fukushima Nuclear Power Plant in 2011, the Covid-19 pandemic started in 2020, and the onset of the full-scale war in Ukraine from 2022.

2 Theoretical Background

It is commonly accepted that a concept is observable in discourse through its verbalisers, among which we can distinguish the core lexeme and peripheral markers. In a survey, Lőrincz (2022) outlines perspectives for exploring themes of attitudes towards war and its influence on the global economy on the basis of a corpus representing the full-scale war in Ukraine by applying a set of subject keywords and their collocates: "Direct negative representation illustrated the war in Ukraine and the general negative attitude of the public, as shown based on the analysis of collocates of war" (p. 66).

The idea proves fruitful and is successfully used in the reverse process, i.e., corpus creation on the basis of keywords on the Sketch Engine platform (the so-called *seed words*) (Sketch Engine, 2024). Whereas some themes may indeed be expressed through direct denominations, some relevant themes may be hidden and better observed through indirect representations. As Sun points out, "Emotions can be reflected through direct and indirect linguistic markers. People can directly reveal how they are feeling using positive and negative emotion words (e.g., 'I'm feeling happy', 'I'm so angry'). Language can also carry indirect traces of a person's emotion experience" (Sun et al., 2020). Wierzbicka (1992) concludes that defining an emotion relies on a cognitive scenario composed of thoughts, wants, and feelings (p. 557), which manifests as contextual emotional markers. Vilchynska (2018) emphasises the role of observing associative and semantic relations to ensure concept integrity: "Lexical meaning inherent to its verbalisers, its distribution, word-formative valence, the possibility of reviving of the ancient associative and semantic relations of certain words ensure the integrity of considered sacral ethnoconcept" (p. 116). For instance, by analyzing

co-occurrence patterns, Mizin and Slavova (2023) demonstrate that "[the] lexeme *jealousy* conveys in English a wider range of emotional experiences, because it contains semantics of not only jealousy (romantic envy), but also intolerance or hostility to the superiority of others, i.e. it partially coincides with the meaning of the lexeme *envy*" (p. 2). Logically, associative relations should include both direct and indirect concept verbalisers.

The adjective "indirect" used in terminology may encompass a broad meaning due to its figurative nature. One of the efficient linguistic devices to create associative relations among concepts is metonymy. Therefore, the core concept and a considerable part of its indirect markers are linked through a sort of metonymic semantic relation. This idea does not appear to be a brand-new statement; however, it seems scarcely explored. Vu dwells on its potentiality in hidden-sense creation: "conceptual metonymy obviously contributes significantly to the formation of implicit meanings of idioms" (Vu et al., 2020, p. 30). For example, Kövecses highlights that "the major conventionally verbalised conceptual metonymies for anger in English include body heat, internal pressure, agitation, and interference with accurate perception" (Kövecses, 2004).

As Stefanowitsch emphasises regarding the concept of FEAR, "the most salient aspect of the English concept 'fear' does not seem to be the experience of the emotion itself, but the consequences of that experience (cf. section 2.3)" (as cited in Oster, 2012, p. 334). This is consistent with the view that fear is a multifaceted concept, classifiable by its source, level of control, duration, intensity, and timing (Mykhalchuk & Bihunova, 2019, p. 4). Subsequently, we hypothesise that the traits of the emotion of FEAR may be indirectly expressed through things used to prevent dangerous situations and, logically, appear in a marker of DISEASE, since it is associated with the prevention of catastrophic consequences for health. Thus, in this survey, we intend to subject an indirect lexical marker of the discourse of catastrophe, such as DISEASE and WAR.

3 Methodology

In many domains, researchers need to make judgements about unobservable processes on the basis of indirectly related observable features. In linguistic surveys, formal parameters can allow for the inference of certain hidden extralinguistic features. One of these features is the conceptualisation of reality in one's mind, which may be codified in a text or discourse and can be brought to light through the observation of its formal traits:

> Thus, the extralinguistic reality is not an "intangible" abstract entity but it can have a quite material projection in the source and target texts due to the presence of specific Indicators of Extralinguistic Reality (IREs) available for perception (though not always consciously) by the addressee, translator and researcher. The least studied IREs are those implemented at the macro level (of a text or a set of texts). These indicators can be segmental and suprasegmental, linear and non-linear, conscious and unconscious, indicating the author's personality, the author's intention, traditions, or other aspects of the culture to which the author belongs, etc.[1] (Fokin, 2013, p. 133)

Concept verbalisers, along with their distribution frequency, seem to be good candidates for these indicators. Vassey demonstrates that the exploration of word frequency can shed light on the ideology of media news: "the comparison of metalanguage across corpora – and even corpora of different languages – can also help researchers explore (potentially ideological) representations of languages within data sets drawn from different social groups" (Vessey, 2017, p. 279). In this respect, Kemppanen (2004) concludes that the concept of FRIENDSHIP in Finnish original texts and texts translated from Russian is used differently in the original and in the translation due to ideological factors. Among the observations regarding the number of translation shifts, frequency of lexemes used includes also instances of omissions; for example, the omission of the word *occupation* in translation may testify to the influence of a translator's ideology (Gumul, 2011).

[1] *Our translation*.

Romanyshyn (2020) examines key concept verbalisers related to Ukraine based on their absolute and relative frequencies. Berber Sardinha (2002) uses statistical tools to evaluate the effectiveness of automatic keyword extraction depending on the reference corpus volume. Given volume constraints, we cannot provide an in-depth overview of many valuable contributions by different authors exploring the frequency factor in concept verbaliser usage. At the same time, it is evident that numerous surveys highlighting either the conceptual, stylistic, genre, or technical aspects – such as the software used, statistically significant test methods, or corpora creation and processing – are dispersed across different works, papers, or manuals. Thus, in this survey, to assess the evaluative level of the concepts in question, attention was paid to the frequency and distribution of the lexeme *iodine/йод*, which may carry evaluative and emotional meanings in two language communities – English-speaking (USA) and Ukrainian-speaking (Ukraine) – as a verbaliser of the concepts DISEASE and WAR.

We employ a case study to explore whether the yearly usage of the noun *iodine* can be associated with particular events and to examine whether this usage is statistically significant, emotionally relevant, and, logically, classifiable as a concept. The survey was conducted on data extracted from two corpora: the NOW Corpus (Davies, 2024), currently comprising 4,904,641,315 tokens, and the GRAK corpus (General Regionally Annotated Corpus of Ukrainian), which contains 1,781,522,761 tokens. The first corpus is part of a broad corpora collection created by Mark Davies; therefore, the tools described here are also applicable to other English, Portuguese, and Spanish corpora created by this scholar. The GRAK corpus runs on the Sketch Engine platform, which currently integrates over 700 ready-made corpora in more than 100 languages, and the described sampling methodology for this corpus is also reusable for other Sketch Engine corpora.

3.1 Sampling Methodology

We primarily obtain a dataset mapping the absolute frequency of the lexeme *iodine* to the corresponding years from English and Ukrainian corpora. To do this, we first enter the term in question into the search field of the NOW Corpus (Davies, 2024). On this platform, it is necessary to capitalise the searched lexeme in order to generate concordances for the entire paradigm. However, since *iodine* is uncountable, extracting singular number forms is sufficient; therefore, the query is performed using lowercase letters.

To limit the selection to a single country, the "United States" option is selected within the "Sections" drop-down menu. The frequency data for the English (USA) dataset was then extracted by utilizing the "Chart" function and filtering results by year.

The generated table can be copied into a spreadsheet program, such as Microsoft Excel or Google Sheets. To ensure data integrity during processing, the horizontal table is transposed vertically using the `TRANSPOSE` function, and duplicate entries are removed to obtain a clean dataset of unique yearly occurrences.

The result of the pasting is shown in Table 1.

Subsequently, it is necessary to obtain a corresponding statistical dataset for the Ukrainian language from the GRAK corpus (version 17) regarding the absolute frequency of the lexeme *йод*. Although the resulting table is similar, the procedure differs because the corpus runs on the Sketch Engine platform, which uses a different query language called Corpus Query Language (Sketch Engine, n.d.). There are several ways to retrieve statistical data from a Sketch Engine corpus by year; however, we intend to illustrate one possible option for data extraction.

Once the simple query *йод* is entered into the search form, we obtain concordances with a total of 4,959 occurrences of this lexeme. To group the results by year, it is advisable to download the entire table by pressing the "Download" button and choosing the "xls" format. Following the extraction, the year is separated from other reference data using the `=LEFT(A2, 4)` formula, allowing for the automation of the grouping process by year through the `COUNTIF` function. This procedure yields the absolute frequency of the lexeme *йод* in the Ukrainian corpus, as presented in Table 2.

The data from this table is still insufficient to draw conclusions. Although there appears to be a general upward trend in the absolute frequency, this may only be apparent since the amount of available digital text is also growing from year to year. Thus, to measure the real frequency of the lexeme *йод*, it is necessary to consider its relative frequency. The relative frequency of a phenomenon is the ratio of the absolute frequency

Table 1. Absolute and relative frequency of the lexeme *iodine* per year.

Year	Freq.	Words (M)	Per mil
ALL	3960	18400	0.22
2010	72	84.7	0.85
2011	370	87.6	4.23
2012	146	90.4	1.61
2013	115	84.9	1.35
2014	80	82.7	0.97
2015	99	92.5	1.07
2016	91	169.4	0.54
2017	50	200.5	0.25
2018	115	185.4	0.62
2019	416	486.4	0.86
2020	783	1138.9	0.69
2021	703	1278.2	0.55
2022	652	1032.9	0.63

Table 2. Absolute frequency of the lexeme *йод* per year.

Year	Absolute frequency of use of the lexeme *iodine*
2010	203
2011	354
2012	264
2013	249
2014	135
2015	223
2016	138
2017	168
2018	285
2019	282
2020	342
2021	213
2022	366

to the total volume of data. By clicking the "I" (i.e., "Information") button near the corpus version and downloading the "doc. Date" option, the volume for each year in the Sketch Engine corpus is displayed (see Figure 1).

The relative frequency, obtained by dividing the absolute values by the respective subcorpora volumes, can be calculated either manually or automatically (see Table 3).

3.2 Statistical Methodology and Discussion

In Table 4 and Figure 2, we can observe that the relative-frequency fluctuations for both datasets mostly correlate: usage drops in 2012, 2014, 2016, and 2021 in both corpora, whereas in 2020, this only occurs in the USA subcorpus dataset.

To objectively verify whether the datasets indeed correlate, and to ensure that similarities in the relative frequency changes across years in the USA and in Ukraine are not due to chance, several statistical tools may be used. One of the most common is the Pearson correlation coefficient, which quantifies the degree of linear relationship between two continuous variables. In other words, it assesses how well a straight-line relationship describes the relationship between the variables obtained from both datasets.

To calculate the Pearson correlation automatically, a simple formula is used, with both dataset ranges specified inside parentheses and separated by a comma: `=PEARSON(B2:B14, C2:C14)`.

For the obtained data, the Pearson correlation equals 0.65067, which is classified as high (DATAtab Team, 2024). This means that incremental and decremental trends in the two datasets are indeed interconnected. Observing the data, we can see that in 2022 the relative frequency of this lexeme in Ukrainian

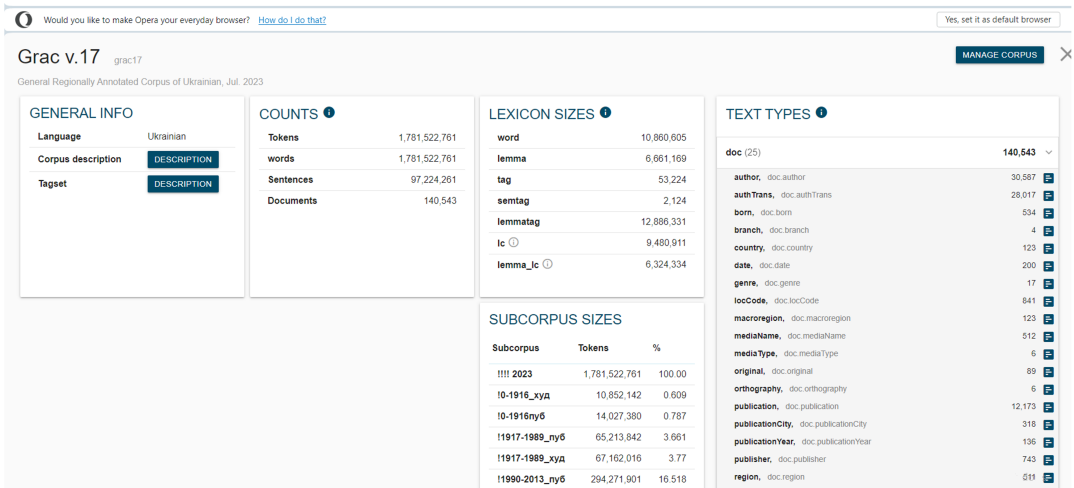


Figure 1. Accessing the volume data of the subcorpora.

Table 3. Absolute and relative frequency of the lexeme *йод* per year.

Year	Absolute frequency (per million tokens)	Relative frequency (per million tokens)
2010	203	3.60
2011	354	5.48
2012	264	3.74
2013	249	4.42
2014	135	1.82
2015	223	2.35
2016	138	1.30
2017	168	1.51
2018	285	2.32
2019	282	2.41
2020	342	2.76
2021	213	2.31
2022	366	5.16

more than doubled, whereas in English it increased by only 14.5%. This anomaly could be explained by the full-scale war of Russia against Ukraine, which began on February 24, 2022; this event affected most texts produced during that year, and the exceptionally high usage of the lexeme *йод* might be due to the fear of a nuclear strike in Ukraine, thus being reflected in the Ukrainian GRAK corpus. Iodine is known to be a prophylactic agent against certain types of radioactive intoxication.

It is reasonable to assume that prior to 2022, the correlation of usage in both countries would have been higher. Indeed, if we recalculate the Pearson correlation for the range of years 2010–2021 (rows 2–13 in our case), it results in 0.822631, which is ranked as very high (DATAtab Team, 2024). Let us also observe that in 2020, there was a noticeable increment in the usage of the concept in question in both corpora, followed by a reduction in 2021; this was likely due to the COVID-19 pandemic and the role of iodine as a potential prophylaxis for pulmonary infection.

This example illustrates a general principle: statistical tools do not prove a fact themselves; rather, they show that the observed correlation in the collected data is not merely coincidental. Distinguishing correlation from causation, however, requires human interpretation. Until other possible hidden factors are identified, we may reasonably assume that the COVID-19 pandemic and the full-scale war of Russia against Ukraine caused the observed correlation, expressed in high or very high Pearson correlation values.

Table 4. Correlation between the usage of the lexeme *iodine* (USA) and the lexeme *йод* (Ukraine) across years.

Year	Normalized frequency *iodine* (USA)	Normalized frequency *йод* (Ukraine)
2010	0.85	3.60
2011	4.23	5.48
2012	1.61	3.74
2013	1.35	4.42
2014	0.97	1.82
2015	1.07	2.35
2016	0.54	1.30
2017	0.25	1.51
2018	0.62	2.32
2019	0.86	2.41
2020	0.69	2.76
2021	0.55	2.31
2022	0.63	5.16

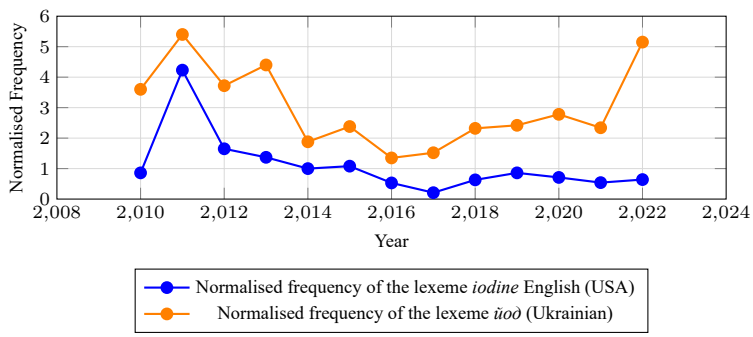


Figure 2. Comparison of normalised frequencies for English and Ukrainian datasets.

A potential drawback of the Pearson correlation is its linear simplification of the data, meaning that the data is processed as the best approximation by a linear regression. In the considered case, this simplification works well, since the overall trends in the usage of this lexeme show incremental growth. However, as previously noted, these trends are not exempt from oscillation. For cases where non-linear data correlation is expected, the Spearman correlation enables a deeper exploration. The principle behind the Spearman correlation lies in assigning ranks to all data points in the arrays, as ranks often capture the year-to-year evolution of the data much better than raw values. To calculate the ranks, the following formulas are applied to each respective columns: `=RANK.AVG(A2, A$2:A$13)` and `=RANK.AVG(B2, B$2:B$13)`.

Table 5 presents the overall picture obtained for the ranks.

To obtain the Spearman correlation coefficient, we calculate the correlation of the ranks instead of raw values by applying the following formula: `=CORREL(C$2:C$13, D$2:D$13)`. The comma-separated arguments of the formula consist of both ranges subject to comparison. This correlation equals 0.797203, which is classified as strong according to Spearman's correlation levels (*Spearman's correlation*, n.d.). Although this value is quite high, it is slightly lower than the Pearson correlation obtained previously. Examining the data from 2010, we notice a considerable difference between the Ukrainian and English (USA) frequencies. This may be explained by the H1N1 influenza pandemic and the confinement measures announced in Ukraine at the end of 2009 (which were not implemented in the USA). If we exclude the 2010 data, the correlation appears higher: 0.858797, which is considered very strong according to Spearman's correlation levels (*Spearman's correlation*, n.d.). The sharp increase in frequency in both datasets corresponding to the year 2011 is evidently related to the accident at the Fukushima Daiichi Nuclear Power Plant and the importance of iodine in treating potential radioactive intoxication.

Table 5. Calculated ranks for each observation from the dataset.

Year	Ukrainian (Value)	English (USA) (Value)	Ukrainian (Rank)	English (Rank)
2010	3.60	0.85	4	7
2011	5.48	4.23	1	1
2012	3.74	1.61	3	2
2013	4.42	1.35	2	3
2014	1.82	0.97	10	5
2015	2.35	1.07	7	4
2016	1.30	0.54	12	11
2017	1.51	0.25	11	12
2018	2.32	0.62	8	9
2019	2.41	0.86	6	6
2020	2.76	0.69	5	8
2021	2.31	0.55	9	10
2022	5.16	0.63	1	1

In fact, it might seem that by reducing the dataset line by line, one could obtain an almost ideal correlation. Nevertheless, such a limited sample size may call its reliability into question. We require an additional tool to prove statistical significance (i.e., to ensure that the obtained data is not due to chance); in a dataset consisting of only a few rows, the likelihood of randomness would be extremely high. To prove that the data is reliable and significant, we resort to the t-test.

The t-test, among various other statistical analysis tools, is integrated into the Microsoft Excel Analysis ToolPak. The result, displayed in an additional worksheet, shows that the two-tailed p-value obtained is 4.42×10^{-6} (or 0.00000442), which is far lower than the acceptable threshold of 0.05. This indicates that the probability of the observed correlation being due to chance is extremely low, see Table 6.

Table 6. t-Test: Paired two sample for means.

Metric	Variable 1	Variable 2
Mean	2.835	1.1325
Variance	1.546991	1.089257
Observations	12	12
Pearson Correlation	0.822631	
Hypothesized Mean Difference	0	
df	11	
t Stat	8.336099	
P(T $\leq$ t) one-tail	2.2E-06	
t Critical one-tail	1.795885	
P(T $\leq$ t) two-tail	4.41E-06	
t Critical two-tail	2.200985	

The statistical significance test allows us to reject the initially formulated null hypothesis. Thus, the lexeme *iodine* indeed displays the features of an indirect concept verbaliser for the concepts WAR and DISEASE, as it maintains a metonymic relation with these core concepts. As Selivanova suggests, there are three basic models of semantic derivation (Selivanova, 2006, p. 528): metaphorical, metonymic, and classificational. In line with the conclusions of other scholars, similar mechanisms may be appropriate for concept verbaliser creation. Among other means of concept verbalisation, Starko also includes the etymology of words that convey a specific idea, along with evaluative and figurative associations (Starko, 2004, p. 5). This observation suggests that the mechanisms of verbaliser formation are worth exploring, especially in comparison to the mechanisms of meaning formation within individual lexemes.

4 Conclusions

The interpretation of the results, which highlights the correlation between statistically significant anomalies in the frequency of the lexeme *iodine/ŭoɔ̆* and historical events, provides evidence that it functions not merely as a word, but also as an indirect verbaliser of the concepts DISEASE and WAR. Indirect verbalisers can be related to their core concept through a metonymic relation, provided that *iodine* is used as a protective treatment in case of radioactive exposure or as a remedy for viral respiratory diseases. An increased frequency of such a verbaliser may implicitly indicate the speaker's emotional engagement.

Corpus datasets showing the relative frequencies of the lexeme *iodine/ŭoɔ̆* in English and Ukrainian reveal similar frequency shifts from 2010 to 2021, with objectively correlated distributions across years as confirmed by Pearson and Spearman correlation tests. Although correlation does not imply causation, human reasoning helps identify key historical events as causal factors: comparable fluctuations in 2011 are likely connected to the accident at the Fukushima Nuclear Power Plant, while the upsurge in 2019–2021 reflects the impact of the COVID-19 pandemic on the linguistic communities of the USA and Ukraine. The sharp increase in usage in 2022 appears exclusively in the Ukrainian corpus due to Russia's full-scale invasion. A t-test on the collected data yielded a low p-value (0.0000044) – well below the conventionally accepted threshold of 0.05 – confirming that the observed differences were not due to chance. The evident relationship of the verbaliser *iodine* to the concepts of DISEASE, WAR, or other types of nuclear threat may testify to the presence of an underlying emotion of fear among speakers.

As a direction for future research, verifying the presented hypothesis on corpora of other languages – representing historic periods of other communities, including the current timeframe – appears promising. Additionally, exploring the evolving relationship between a concept and its direct and indirect verbalisers aligns with a broader research perspective.

Abbreviations

GRAK – *Heneral'nyĭ rehional'no anotovanyĭ korpus ukraïns'koï movy* (2023) [General Regionally Annotated Corpus of Ukrainian]

References

Baron, A., Rayson, P., & Archer, D. (2009). Word frequency and key word statistics in historical corpus linguistics. *Anglistik: International Journal of English Studies*, *20*(1), 41–67. https://www.researchgate.net/publication/284670002_Word_frequency_and_key_word_statistics_in_historical_corpus_linguistics

Berber Sardinha, T. (2002). Comparing corpora with WordSmith Tools: How large must the reference corpus be? In *The Workshop on Comparing Corpora* (pp. 7–13). Association for Computational Linguistics. https://doi.org/10.3115/1117729.1117731

DATAtab Team. (2024). *Pearson correlation*. DATAtab. https://datatab.net/tutorial/pearson-correlation

Davies, M. (2024). *NOW corpus (News on the Web)* [Data set]. English Corpora. https://www.english-corpora.org/now/

Fernández Zas, M. (2021). Specialised discourse in medical research articles: A cross-linguistic comparison between English and Galician. *Diacrítica*, *35*(2), 146–171. https://doi.org/10.21814/diacritica.622

Fokin, S. B. (2013). Metodolohichnyĭ potentsial indykatoriv pozamovnoï diĭsnosti dlia teoriï perekladu. *Studia Germanica et Romanica: Inozemni movy*, *10*(3), 126–137.

Garcia, D., Garas, A., & Schweitzer, F. (2012). Positive words carry less information than negative words. *EPJ Data Science*, *1*, Article 3. https://doi.org/10.1140/epjds3

Gumul, E. (2011). Translational shifts of syntactic and lexical markers of ideology: Reporting the Iraqi conflict in Polish reprint press. *Meta*, *56*(4), 758–774. https://doi.org/10.7202/1011251ar

Heneral'nyĭ rehional'no anotovanyĭ korpus ukraïns'koï movy (Version 14) [Language corpus]. (2023). https://parasol.vmguest.uni-jena.de/grac_crystal/#dashboard?corpname=grac14_full

Kemppanen, H. (2004). Keywords and ideology in translated history texts: A corpus-based analysis. *Across Languages and Cultures*, *5*(1), 89–106. https://doi.org/10.1556/Acr.5.2004.1.5

Kövecses, Z. (2004). *Metaphor and emotion*. Maison des Sciences de l'Homme; Cambridge University Press.
Lőrincz, M. (2022). A corpus analysis of the representation of war in Ukraine in the international press. *The Journal of V. N. Karazin Kharkiv National University*, *2022*(96), 61–68. https://doi.org/10.26565/2786-5312-2022-96-08
Mizin, K., & Slavova, L. (2023). The corpus-based methodology of close emotion concepts differentiation: A case study of ENVY and JEALOUSY. *Cognitive Studies | Études cognitives*, *2023*(23), Article 2811. https://doi.org/10.11649/cs.2811
Mykhalchuk, N., & Bihunova, S. (2019). The verbalization of the concept of "fear" in English and Ukrainian phraseological units. *Cognitive Studies | Études cognitives*, *2019*(19), Article 2043. https://doi.org/10.11649/cs.2043
Oster, U. (2012). "Angst" and "fear" in contrast: A corpus-based analysis of emotion concepts. In M. Brdar, I. Raffaelli, & M. Žic Fuchs (Eds.), *Cognitive linguistics between universality and variation* (pp. 327–354). Cambridge Scholars Publishing.
Romanyshyn, N. (2020). Application of corpus technologies in conceptual studies (based on the concept Ukraine actualization in English and Ukrainian political media discourse). In *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020)* (pp. 472–488). CEUR Workshop Proceedings. https://api.semanticscholar.org/CorpusID:218648766
Selivanova, O. O. (2006). *Suchasna linhvistyka: Terminolohichna entsyklopediia*. Dovkillia-K.
Sketch Engine. (n.d.). *CQL guide*. https://www.sketchengine.eu/documentation/cql-basics/
Sketch Engine. (2024). *Create a corpus from the web*. https://www.sketchengine.eu/guide/create-a-corpus-from-the-web/
Spearman's correlation. (n.d.). Statstutor. https://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf
Starko, V. F. (2004). *Kontsept Hra v konteksti slov'ians'kykh i hermans'kykh kul'tur (na materiali ukraïns'koï, rosiĭs'koï, anhliĭs'koï ta nimets'koï mov)* [Summary of doctoral dissertation, O. O. Potebnia Institute of Linguistics of the NAS of Ukraine]. https://uacademic.info/ua/document/0404U001785
Sun, J., Schwartz, H. A., Son, Y., Kern, M. L., & Vazire, S. (2020). The language of well-being: Tracking fluctuations in emotion experience through everyday speech. *Journal of Personality and Social Psychology*, *118*(2), 364–387. https://doi.org/10.1037/pspp0000244
Tissari, H. (2001). Affection, friendship, passion and charity: A history of four "love lexemes" since the fifteenth century. *Neuphilologische Mitteilungen*, *102*(1), 49–76. https://www.jstor.org/stable/43344799
Velasco-Sacristán, M. (2010). Metonymic grounding of ideological metaphors: Evidence from advertising gender metaphors. *Journal of Pragmatics*, *42*(1), 64–96. https://doi.org/10.1016/j.pragma.2009.05.019
Vessey, R. (2017). Corpus approaches to language ideology. *Applied Linguistics*, *38*(3), 277–296. https://doi.org/10.1093/applin/amv023
Vilchynska, T. (2018). Features of verbalization of the sacral concept "God" in the Ukrainian ethnolingual culture. In L. Struhanets (Ed.), *Multidimensionality of vocabulary in the literary language* (pp. 104–116). Lambert Academic Publishing. http://dspace.tnpu.edu.ua/bitstream/123456789/13871/1/struhanets_monogr_Multidimensionality.pdf
Vu, N. N., Van, N. T. T., & Lien, N. T. H. (2020). Cross-linguistic analysis of metonymic conceptualization of personality in English and Vietnamese idioms containing "head", "face" and "eyes". *International Journal of English Language Studies*, *2*(4), 23–32. https://doi.org/10.32996/ijels.2020.2.4.3
Wierzbicka, A. (1992). Defining emotion concepts. *Cognitive Science*, *16*(4), 539–581. https://doi.org/10.1016/0364-0213(92)90031-O

The publication was financed at the author's expense.
This publication was reviewed using the CLARIN-PL research infrastructure (http://clarin-pl.eu).
The author declares that he has no competing interests.

This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0), which permits redistribution, commercial and non-commercial, provided that the article is properly cited.

© The Author 2026

Publisher: Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland
Publishing history: Received 2024-07-16; Accepted 2026-01-30; Published 2026-11-13.