RAEL: Revista Electrónica de Lingüística Aplicada
Vol./Núm.: 21/1
Enero-diciembre 2022
Páginas: 1-18
Artículo recibido: 09/05/2022
Artículo aceptado: 30/11/2022
Artículo publicado 31/01/2023
Url: https://rael.aesla.org.es/index.php/RAEL/article/view/487
DOI: https //doi.org/10.58859/rael.v21i1.487

Dynamic Synoptic Scheme as a Tool for Searching Phraseological Synonyms

Esquema sinóptico dinámico como herramienta de búsqueda de sinónimos fraseológicos

SERHII FOKIN
TARAS SHEVCHENKO NATIONAL UNIVERSITY OF KYIV

The typical modern-day thesauri cover the lexeme level, whereas clichés, aphorisms, and phraseology are underrepresented in their entries. One of the best ways of ordering phrase-level synonyms is by relating them to the keyword descriptors within an onomasiological dictionary macrostructure, which would provide easier access to users. The goal of this article is to describe a methodology of organising a reasonable and handy synoptic scheme for accessing phraseological units in an electronic onomasiological-phraseological dictionary. In the computational format, onomasiological dictionaries become much more manageable, thanks to the availability of the formal search. Since these dictionaries are based upon classificatory synoptic schemes, their compilation requires deep analytical data processing and intellectual effort. An extremely detailed hierarchy would be difficult to use, and, conversely, a simple one hardly covers the most relevant semantic features of the entries.

Keywords: *phraseological synonym, onomasiological dictionary, computational lexicography, synoptic scheme, translation*.

La mayoría de los diccionarios de sinónimos se centran en el nivel de lexemas separados mientras que los clichés, frases hechas y modismos aparecen raramente entre sus entradas. Una de las mejores formas de organizar los sinónimos fraseológicos se consigue gracias al uso de palabras clave que describen su significado dentro de la macroestructura de un diccionario onomasiológico, lo cual ha de simplificar el acceso para el usuario. El objetivo del artículo es describir la metodología de la organización de un esquema sinóptico fácilmente manejable de unidades fraseológicas para un diccionario onomasiológico fraseológico computacional. Ya que dichos diccionarios se basan en un sistema sinóptico clasificatorio, la compilación de los diccionarios onomasiológicos requiere un procesamiento analítico de datos y esfuerzos intelectuales. Una jerarquía muy ramificada resultaría difícil de usar, y, por el contrario, un esquema simplista apenas cubriría las características semánticas básicas de las entradas.

Palabras clave: *sinónimo fraseológico, diccionario onomasiológico, lexicografía computacional, descriptor semántico, esquema sinóptico.*

1. INTRODUCTION

Today’s electronic dictionaries go far beyond mere scanned copies of their paper predecessors. Their functionality is impressive, provided that they are easy to use and fitted with a wide range of new features. For example, the alphabetical search, which turns out to be a time-consuming procedure for poorly experienced users, is notably simplified, thanks to the automatic search feature. This type of search can now be performed across many dictionary modules at a time.

However, while the alphabetical search is still manageable in paper dictionaries, the reverse search (that is, searching for a lexeme by its definition) in a classic paper dictionary would literally amount to looking for a needle in the entire haystack of thousands of entries. An awareness of this issue brought about multiple attempts of creating onomasiological dictionaries meant to search words through their ideas. Despite the vast number of dictionaries that prove successful in this domain, there is still a remarkable gap of knowledge around their usage and compilation principles.

The simplest way of creating an onomasiological dictionary is by inverting the parts of each entry (i.e., by putting the glossa before the lemma). On the other hand, using reverse dictionaries is not as easy as it might seem. For example, if a potential user were to search for all of the entries containing the lexeme *alphabet*, their search in *One Look Reverse Dictionary* would yield over one hundred results, among them *spelling*, *alphabetical*, *orthography*, *language*, etc., which is nevertheless much easier to handle than the whole set of the entries eventually containing hundreds of thousand lemmas. Another classic example of reverse dictionaries, which also allows finding lexemes by their meanings, is *IEDRA* (*Buscador de palabras*). Since their function consists in finding words (or “names”) for a certain meaning, the reverse dictionaries are also considered as a subtype of onomasiological dictionaries (*ὄνομα* means *name* in Greek), although the latter concept is somewhat broader than the former, as is explained *infra*.

While searching for lexemes during text writing or editing can be facilitated with the use of reverse dictionaries, handling sequences of words, set phrases, phraseological units (PUs), etc. could become a much more challenging task with unpredictable results, since many phraseological units do not appear in bilingual dictionaries. What is more, even if they do, their translations often possess different core semantics or connotation from the original one. Logically, an appropriate phrase search can become a time-consuming process, and a translator, text writer, or a journalist would sometimes need to invest hours trying to establish the bridge between the meaning the user had in mind and the phrases existing in the target language. Given these challenges, compiling phraseological onomasiological dictionaries, with comfortable and easily understandable keyword descriptors acting as entry points for the user, can substantially optimize searching for a phrase with a given meaning. They should prove helpful when a writer or translator does not know the necessary phrase in the target language, has it on the tip of his or their tongue, or merely does not remember it, which frequently happens in creative writing.

Some existing multilingual lexicographic tools, such as *AUTOFRAS* (Pamies, Iñiesta Mena, Bálmacz & Káloustova, 1998a) published at the University of Granada (Spain), and *Refranero Multilingüe* (1997-2021) developed by the Cervantes Institute, are of great help for the writers and translators. The latter provides a detailed set of keywords by which a paroemia can be found. For example, the search by the idea *Dinero* yields the following results: *Al buen pagador no le duelen prendas*, *Amor y dinero nunca fueron compañeros*, *El dinero hace caballero, La mujer y el oro lo pueden todo*. In contrast, the search based on the idea *Riqueza* produces quite a different outcome: *Con el buen pan y buen vino, no te faltarán* *amigos*; *De mí te reirás, pero de mi dinero, no*; *Dinero llama dinero*; *El dinero hace caballero*.

As we can observe in both lists of phrases, the only paroemia reached by both ideas, *Dinero* and *Riqueza* is *El dinero hace caballero*. At the same time, the paroemia *Poderoso caballero es don Dinero* is assigned only the idea *Poder* (i.e., *power*). Subsequently, in the case of using multi-label classification (i.e., assigning multiple descriptors to this entry and, in addition to the mentioned ideas: *money*, *richness*, *power*, *authority*, *expenses*, *earnings*, and *respect*) the likelihood of these phrases being found by the user would increase by seven-fold when compared with the probability of these phrases of being found by the user through a keyword search. This approach, however, is not without its drawbacks. The use of multi-label classification would significantly increase the number of keywords, while simultaneously making them more difficult to manage from the standpoint of the program's backend, as well as from the user interface.

Recognizing that the phraseology level is far broader than the paroemic one, we propose and further subject to an experiment the following basic ideas to incorporate into a phraseological dictionary macrostructure: 1) using as many as possible descriptors for each phraseological entry; 2) optimizing the descriptors handling by organizing them in a two-level scheme of categories and subcategories; and 3) applying an algorithm of a dynamic macrostructure generation so the dictionary compiler does not need to worry about recreating the macrostructure when the dictionary is updated with some new entries.

Thus, the goal of this article is to describe a possible methodology of organising a reasonable and handy synoptic scheme for accessing PUs in an electronic, onomasiological-phraseological dictionary.

2. THEORETICAL BACKGROUND

2.1 *Phraseological synonymy scope.*

As stated earlier, bilingual or multilingual phraseological dictionaries do not necessarily provide equivalent translations for phraseological units in terms of semantics, basic motivating structure (or images), connotation, or implications in the context that the phrase is being used at that time. If translational equivalents are sometimes far from that which is desired, do the same stumbling blocks apply to *phraseological synonyms* within the same language? And, given such a disparity in criteria, how should *phraseological synonyms* be defined? How different are they from conventional lexical synonyms? Although interest in this concept has exploded over the last decade, some infrequent definitions can be found before. According to Kunin, *phraseological synonyms* are coreferential phraseological units belonging to the same grammar class, either partially coinciding or completely independent of each other in their lexical structure, holding both common and differential components, either coinciding or differing in their stylistic features (1996: 68). Although the concept of *phraseological synonyms* seems to have crystalized by this time, its practical value for translations was definitely taken into account earlier by Gatiatullina, which is why the researcher introduces the concept of *interlinguistic phraseological synonym* as “PU that coincide by morphologic composition of significant components, by the type of grammatical structure, by common meaning of entire PU in general, but lack inter-linguistic lexical invariant” (1968: 16, as cited in Fazlyeva, 2015).

It is evident that the practical need of using phraseological synonyms may arise both in translation and monolingual communication. But it is in the domain of translation where the issue becomes readily apparent, whereas in monolingual writing or speaking the communicators are less likely to become aware of the need of a phraseological synonym, unless they are seasoned professionals.

Both cited scholars emphasize the need to preserve the grammatical features in a synonym. It might seem commonly accepted that synonyms should belong to the same grammar class. However, once the practical needs of translation are examined, and given that the phraseological level is highly subject to transformations in the target text, one realizes that the transpositions, modulations and other procedures, widely used in translation, can often cause the grammatical structure to change in the translated text. Subsequently, two Spanish PUs *darle un pronto* (verbal phrase) and *de buenas a primeras* (adverbial phrase) could be interchangeable in a sentence after some necessary syntactic adaptations. This may help explain why this former verbal phrase appears translated as an adverbial phrase *without warning*, *suddenly* (Como se dice en, 2013). At the same time, it often proves difficult to find more suitable translations for the indicated phrase (such as *all out of the blue* or similar) in open sources accessible through a search engine such as Google. However, this example (as with many other possible illustrations of this kind) allows us to ignore the grammatical structure as a requirement to rule out phraseological synonyms.

Additional issues on phraseological synonyms definitions have been examined by Rodríguez-Piñero, who queries whether phraseological synonyms could be variants of the same PUs, and whether PUs with different distribution, and semantic combination should be considered as phraseological synonyms (2012: 235-236). For their part, Dobrovol'skij and Baranov observe that images disparity in coreferential phrases raises questions concerning their synonymy; moreover, the issue of *quasisynonymy* should be defined in the domain of phraseology as well (2011). With respect to the latter, an example is provided by Mellado Blanco (2014), where *ir al grano* y *hablar sin rodeos* share the same meaning. However, the former is not limited to speaking exclusively, which is why they are not interchangeable in some contexts and thus do not fit neatly into the concept of a quasi-synonym. Therefore, Piñero concludes, that a difficulty faced by researchers is the lack of criteria to classify some PUs as such (just collocational or fixed phrases), their different syntagmatic combinatory, pertinence to different domains of usage, as well as their polysemy (2011: 22).

In addition to *phraseological synonyms*, Fazlyeva extends the types of phraseological similarity relations with *phraseological analogues* (2015), a concept allowing somewhat wider freedom in the interpretation than 'phraseological synonym':

> We include phraseological analogues to the type of interlinguistic relations. Analogues are understood to be set expressions, which are adequate by meaning of original language PU, but completely or partially differ from it by image. Analogues impart specific, different images or notion of different nations. One and the same reality can be delivered with different lexemes in different languages; differences may affect structural-grammatical organization of PU. It’s impossible to describe appearance of analogues only in semantic terms (Fazlyeva, 2015: 7).

If we focus on the practical need, beyond strictly descriptive objectives, it is worth noting that the need for translation may allow changing either the grammatical structure, lexical components, images, or even semantic features, in order to render the desired invariant. Finally, if two PUs can be substituted for one another in some contexts, even despite being quasi-synonyms, they are worth being included in a dictionary of phraseological synonyms, since it is the first reference point from which a user might decide to search for them. Given these assumptions, it seems reasonable to frame the concept of phraseological synonym as broadly as possibly so that users may have at their choice as many variants as possible with the advantage of filtering the results in case the query yields many phrases. Once a list of synonyms is found in a dictionary, a set of new issues arises regarding their correct usage. As close as two PUs might be, a user might not be aware of some additional connotations or implications a PU may hold. Logically, the possibility of usages in the given or similar context the speaker has in mind matters the most, rather than the grammatical structure, images or other features. At this point, this is the only criterion to state whether two PUs are synonyms or not. A dictionary compiler cannot foresee all the possible contexts; thus, there is no sense in applying the criterion of the context for a dictionary of phraseological synonyms. However, the user could handle this constraint on their end, which is why observing the phrase usage in a concordance is highly recommended. *IdeoPhrase* is provided with a feature allowing to look up the usage of the phrase in question *in corpora* and to decide whether the PU is suitable for the context of speaking.

2.2 *Conventional approaches to the onomasiological dictionaries' macrostructure*

The first attempts at organising dictionaries according to thematic categories through classificatory schemes (also called "synoptic schemes") date back to ancient times. In 170 A. D., Pollux compiled his famous *Onomasticon*, a dictionary with features of an encyclopaedia, subdivided into ten thematic books. In the Middle Ages (around A.D. 750), Aban Ibn Taghlib compiled Kitab covering a set of subjects in an analogous way. Roget published his famous *Thesaurus of English Words and Phrases* in 1852, which consists of six books (2011). More concretely, the Thesaurus divided 1,000 concepts into six sections. In this domain, Casares reached a significant milestone thanks to his *Diccionario Ideológico de la lengua española* designed in 1942 (1994). While Pollux's work was grounded on a one-level and straightforward classification, Casares's scheme comprised on average up to eight (and, in some subsections, up to 11) nested levels. For example, one of the longest routes possible in the entanglement of lexical system, starting from the widest category and moving down to more and more narrow domains, might be the following: *El Universo* - *Mundo orgánico* - *Reino animal* - *El hombre* - *El individuo* - *El individuo como sujeto racional* - *Inteligencia* - *El conocimiento a priori* - *Espacio* - *Movimieno* - *Adelantamiento* - *Choque* (1994: XXXVI). The last link, for example, redirects the potential user to a list of related lexemes: "choque, impacto, impacción, encuentro, encontrón, encontronazo, estrellón, reencuentro, topetón, topetazo, tope, colisión, abordaje, trompada, beso, pechugón, trompicón, trompilladura, tropiezo, tropezón, tropezadura, traspié, cambalud, trastabillón" (Casares, 1994: 120). One can see that the choice between these two extreme approaches could be a trade-off between an extreme granularity and just a plain synoptic scheme. As Popović states, the hierarchization is important with regard to precision in cases "when there are many idioms under the entry" (2020: 147).

In the domain of computational technologies, a new range of possibilities for lexicography arise, such as interactivity (users may actively contribute to correcting or adding entries), or the possibility of reorganising macrostructure on the fly by performing search throughout different structural parts of a dictionary (not only by lemmas). And it is through these possibilities that we find the idea of organising a dynamic synoptic scheme instead of a static one.

Although modern electronic onomasiological dictionaries are seldom made on the basis of synoptic schemes, they are generally inverted-structured search engines, where users can find a lexeme by introducing (fully or partially) its definition. It seems strange that many dictionaries do not have such schemes, considering the vast availability of open-source databases and particular software for flipping fields in a given dataset. Among dictionaries that offer inverted search, we should also mention *SUM* (i.e., *Academic Explanatory Dictionary of the Ukrainian Language*, 2018*)*, the *Dicionário Priberam da Língua Portuguesa* (2019), the already mentioned *Diccionario Inverso de la Real Academia Española*, and *OneLook Reverse Dictionary* of the English language. The function of the inverted search was developed as an interesting collateral feature, easily attainable due to technological progress rather than a response to an objective demand. At the same time, some rare pearls expressively elaborated for onomasiological usage are little known, and remain largely disapproved by the philological community due to a lack of knowledge as to their usage and possibilities. Taljard and Prinsloo describe the relevant issue in a pictographic dictionary for children, where the lemmas are accessible through pictures representing the concept to be searched for:

> In the case of children's dictionaries, it is almost inevitable that the (adult) lexicographer's frame of reference and that of the child as target user will not coincide, resulting in an unsuccessful transfer of the lexicographic message to the user, or simply put, an unsuccessful dictionary consultation (2019: 208).

It can be said that electronic onomasiological dictionaries (also alluded to as ideological dictionaries) are classifiable in two groups: those which are based on a simple inverted structure (i.e., placing glossa before lemma), and those which are formulated around a hierarchically organised synoptic scheme. Thanks to this kind of scheme, the lemmas are accessible under a certain general category or a specific subcategory, which renders the search much more precise and accurate.

2.3 *Synoptic schemes' structure in electronic dictionaries*

The usage of such schemes, which constitutes a revolutionary breakthrough in the art of lexicography, is nevertheless quite difficult from the point of view of intellectual efforts on the user's side: formulating the appropriate keywords or descriptors, trying different options to achieve the desirable results implies deep linguistic intuition and vast general erudition. When we need to find a way of expressing an idea, we can hardly figure out the exact definition. Fortunately, the computational technique can substantially contribute to facilitating both the structuring of this kind of dictionary and the ability to search throughout its database.

Let us assume, for example, that we are searching for a term that denotes "interpreting performed at the same time as speaking". In *OneLook Reverse Dictionary* we do find this term, although it appears in the search results at the 48th position, while on Google's search engine the very first result fitting this description contains the necessary term (i.e., *simultaneous interpreting*). Nonetheless, in the case of some abstract terms with subtle sentimental background, as, for example, *nostalgia*, the exact description turns out to be even trickier. Paraphrases such as 'when one misses one's past' or 'when you miss your past', 'long for your past' are circumlocutions, which, however naive they may appear, do not yield any decent results among the first mentions either on *Google*, or in *OneLook Reverse Dictionary*. In both cases the users do not need to adjust their query to the exact dictionary definitions; the search results will be achieved by introducing some keywords only. The sentence with corrections is already placed instead of the previous one. The latter example also highlights the seriousness of the issue, although, the problem of conceptualization in describing the meaning is beyond the scope of this article, and is exhaustively treated in Sierra's article "Natural language searching in onomasiological dictionaries" (2008). Let us turn our attention to the ways of providing access for a conceptualized form once formulated by a user.

With regard to the previous analysis, it is clear that a mere intuitive user's input is not deemed to be a highly satisfying solution. Some mechanisms of accessing lexicographic units should be well elaborated and prompted to the user, although it appears to be a serious methodological problem. Schryver points out in this respect:

> At the start of the 1990s, databases with ever-growing storage capacities led to dreams in which databases would one day combine alphabetically and thematically ordered dictionaries in one. By the end of that same decade, thematically structured search paths have indeed been developed in addition to the better-known alphabetical search path, especially for electronic encyclopaedias. Nonetheless, the approach to onomasiological EDs has started to shift from a mere focus on database size to clever search mechanisms by which traditional alphabetically organised dictionaries are searched from within an article to the lemma sign (#66). Here an inventive use of specific search words, labels, boolean operators, article fields to be searched, etc., can for instance lead from a combination of keywords in a definition to the item(s) one is looking for. (2003: 175).

For his part, Sierra provides description of componential analysis allowing the user to retrieve descriptors for entries structuring, but even the most complete description of a concept can lack “essential” properties from a user’s point of view. None of the methods of componential analysis, even the most open ones, have been sufficient to foresee the properties used by a small set of students. That gap should be filled with the aid of a good onomasiological retrieval system. This is not to suggest that we will be unable to design a complete and efficient onomasiological dictionary (2008: 37). Another promising approach consists in using word association norms, which an ordinary user without specific background on semantics could apply to obtain results, “although the graph built with all the nodes and edges contained in the datasets tends to be unreliable, due to the number of paths that lead to the wrong results” (Reyes-Magaña, Bel-Enguix, Sierra & Gómez-Adorno, 2019: 887).

Another way of organising the macrostructure of an onomasiological dictionary is by grouping lexemes in accordance with synonymic nodes, each of which is indexed alphabetically, and Schryver mentions *WordNet* as a unique implementation in this regard (2003: 175). We would take the liberty of adding a modest proposal to this list, particularly with respect to the electronic onomasiological-phraseological dictionaries. First of all, we would like to mention several onomasiological-phraseological dictionaries along with their synoptically determined macro-structural features which are scarcely known nor cited in modern surveys.

Multilingual Electronic Phraseological Dictionary *AUTOFRAS* (Pamies et al., 1998a) integrates PUs in 10 languages grouped by common semantic descriptors; the dictionary combines both onomasiological and semasiological features, as it is evident by the dictionary macrostructure. Other rare findings, under-regarded by the philological community, are Ukrainian onomasiological-phraseological dictionaries, the so-called “dictionaries of phraseological synonyms”, so their names are more understandable for ordinary users: *Dictionary of Phraseological Synonyms* (Slovnyk) at web-portal www.rozum.ua and *Dictionary of Phraseological Synonyms* (Dictionary of phraseological synonyms) by Kolomiiets and Rehushevskii (1988), although in paper format, is perfectly suitable for being converted into a digital representation. The dictionary is organised in entries, and ordered by descriptors of ideas; each entry contains phrases corresponding to a certain idea.

Nevertheless, most electronic dictionaries (either explanatory or thesauri) are focused mainly on lexemes represented by separate words, while users may also need to find synonyms of units far wider than a word: a phrase, cliché, PU, or even sentence. More problems may arise when we need to express a certain idea in a phraseological way, either for a more vivid expression or to put it more plainly. For example, the idea 'to have a dilemma' could be also expressed by means of classic aphoristic phrases such as *Gordian knot*, *between a rock and a hard place*, *Buridan’s ass*. Nevertheless, it is difficult to access these phrases in actual dictionaries of synonyms for two reasons: 1) there are a variety of ways to express the idea, which may be covered by such nouns and verbs as *indecision*, *vacillation*, *dither*, *hesitation*, *hover*; and 2) since the majority of actual electronic dictionaries of synonyms generate one-word synonymy, we very seldom encounter phraseological synonyms within them.

2.4 Organizing dictionary entrance points for users

The described issues of organizing the entrance points for an onomasiological dictionary along with predominance of one-word synonymy in the modern dictionaries constitute a loophole that can and should be covered with the aid of computational technologies. In response to this challenge, we decided to create a multilingual dictionary of phraseological synonyms, which was certified in 2018 under the name of the Multilingual Dictionary of Phraseological Synonyms IdeoPhrase (2020). This dictionary is meant to generate a list of synonyms out of a phraseological database, both in the language of the query or in other languages, more concretely, in English, French, Hebrew, Italian, Latin, Russian, Spanish, and Ukrainian. The onomasiological organisation of phraseological dictionary is aimed at partially bridging the gap left by the lack of phrase-level synonyms in modern dictionaries, as well as searching stylistically marked means of expressing a certain idea, since phraseological units are typically a more powerful stylistic engine than their correlative plain, unmarked synonyms.

As it can be observed, the onomasiological approach for compiling phraseological dictionaries is not accidental. As Pamies, Balmacz and Iniesta Mena specify, some translations of PUs are quite approximate, either by their semantics or by connotative features (1998b: 207). We assume that it is more comfortable for a user to have at their disposal a wide choice of semantically related PUs to find those which better fit the extra-linguistic parameters of translation, rather than relying on finding an exact translation, as it should be in the case of terminology. In contrast, an exact formal search of PUs in some cases might be restricted by one of the lexemes, and it is sometimes difficult to establish which one is the main lexeme to be used as the basis of the search. Therefore, the usage of onomasiological dictionaries is not necessarily more difficult than that of semasiological ones in the domain of phraseology (Pamies et al., 1998b: 207).

Modern computational tools include a wide range of approaches that can be used to accomplish the task of classification, owing primarily to the progress in artificial intelligence. A wide array of methods for automatically clustering most variable types of data could be extended to the analysis of several language levels, particularly grammar, vocabulary, and phraseology. On the other hand, despite the evident advantages of artificial intelligence for accomplishing classifying tasks, some features and stages of its performance rest “behind the scenes”, whilst it is often important for the scientist to be aware of each algorithm step and to know the reasons to put a hypothesis to the test, argue for it, and possibly decline it. However, published works in this domain are not very prolific; not surprisingly, most research delves directly into text classification problems. A text possesses an enormous set of discreet, easily retrievable formal features (e.g., words, sentences, paragraphs, and punctuation signs), while a separate lexeme has a quite limited number of significant structural parts and visible features. Logically, its classification involves using descriptors in lexicographic resources (for example, explanatory or translational dictionaries), which are considered neither objective nor complete. Corpora seem to be a more objective piece of data, and in this field the proposal by Zhao of classifying English vocabulary on the basis of lexemes contextual relation is fundamental, objective, and promising (2018). In other words, in the absence of rich, internal data concerning vocabulary, classifying by means of external data is evidently deeper. Meanwhile, there is also an objective need to explore the potential of classifying terms on the basis of inner data, such as dictionary entries.

3. METHODS

The methodology used for organizing the synoptic schema serving as entrance points for the users is based upon observation and classification. Initially, a set of descriptors (tags, keywords) are assigned to each phraseological entry (as per Table 1). These descriptors are assigned first intuitively and, at further stages, in accordance with the previously used descriptors. Once the relational database is complete, the program chooses the most frequent descriptors to use them as keywords. These keywords serve as entrance point at the first hierarchical level for a user. One level scheme may be enough for some purposes and provides access to all the phraseological units by a chosen keyword. To organize the hierarchy into a two-level scheme, i.e., to provide access to the phraseological units through descriptors and subdescriptors, the program takes into account the co-occurrence of the descriptors. For each first-level descriptor, the program chooses those most frequently occurred with them within the same entries. The most frequent co-occurring ones are stored as subdescriptors. As the dictionary gets updated with new phraseological units and their respective descriptors, the program will reorganize the synoptic schema bases upon the frequency and co-occurrence factors. Let us explain hereafter each step in a detailed way.

Our general assumption is based on the idea that a universal synoptic scheme, even if it were possible, would be hardly usable in onomasiological dictionaries aimed at practical purposes. By contrast, a synoptic scheme generated *ad hoc* (i.e., dynamically compiled), could be much more practical and workable, and would better satisfy the criterion of interactivity, whereas immutable static schemes used until now lack such possibility.

We have chosen the database entries as empirical material for the experiment. Since the entries of the dictionary contain semantic descriptors, as it is shown below in Table 1, they serve as a basis for calculating semantic distance and searching phraseological synonyms. The semantic distance here is considered as a magnitude directly proportional to the number of coincident descriptors. We are aware that there is a set of more specific ways of calculating semantic distance, as Arapov's-Ratseva's or Shreider's methods (Skorokhodko, 1970: 181-182), and this choice has been argued by Fokin in his article "Neural network pattern for enhancing the functionality of electronic dictionaries" (2019) as convenient for this particular kind of dictionary. Modern computational lexicographic resources provide researchers with an ever-growing number of specific tools for calculating semantic similarities with more varied purposes. In particular, Cooper describes results of computing semantic distance performed on the basis of comparing structures in bilingual dictionaries entries (2008). Other proposals (Tsang & Stevenson, 2004) of calculating semantic distance are grounded on exploring the database of lexical relations *WordNet* (WordNet), since semantic relations of a lexeme can reflect their semantic properties. For example, Kenett, Levi, Anaki, and Faust point out that semantic distance can be measured by calculating "the amount of steps needed to traverse from one word to another" (2017). A similar approach involving the usage of a predefined word hierarchy which has words, meaning, and relationship with other words stored in a tree-like structure has been explored by Pawar and Mago (2018).

Despite a considerable breakthrough in the domain of lexical semantics in the context of computational lexicography, rare are the works which mention the issue of phraseology, and even fewer are those which are focused on onomasiological dictionaries. In Table 1, we show a snippet from the database containing 7 entries:

Table 1. ***Snippet of the database involved in the experiment***

English	Ukrainian	Spanish	Russian	Hebrew	Descriptors
to know something on good authority	з перших рук	de buena tinta, saber de buena tinta	из первых рук	מיד ראשונה	precision, source, first, truth, correctness
to get to the point, to get down to brass tracks	переходити до справи	ir al grano	Переходить к делу, переходить к сути	תגיע לנקודה	understanding, essence, reason, motive, origin, precision, start
face to face	віч-на-віч	cara a cara	с глазу на глаз	בארבע עיניים	secret, mysteriousness, privacy, privacy, chat
English	**Ukrainian**	**Spanish**	**Russian**	**Hebrew**	**Descriptors**
sharp tongue, viper's tongue	гострий язик	tener una lengua afilada	острый язык	לשון חדה	wit, censoriousness, speech, word, annoyance, criticism, precision, accuracy
to lay cards on the table	відкривати карти	poner las cartas boca arriba	раскрывать карты	הקלפים על השולחן	opening, secret, disclosure, search, trust

The PUs in the far-left column of Table 1 were processed as phraseological entries, i.e., taking for granted their metaphorical meaning, not literal meaning. The descriptors in the last column of Table 1 have been manually assigned to each phraseological entry; first intuitively, as per its semantic and context of usage, and then substituted with the most commonly used in describing multiple entries. It turned out that few descriptors lacked a prolific usage. This procedure was performed according to the Method of Lexicographic Portrait, developed at the Maurice Thorez Moscow State Pedagogical Institute of Foreign Languages: lexicographic components are first selected intuitively, among which pivot categories are segregated (as cited in Darchuk, 2008: 243). Once performed this assignment and having proved the functionality of the dictionary, we came to the conclusion that, as expected, the number of descriptors itself is to be restricted to a certain set (otherwise it would be impossible to establish a common denominator among descriptions of PUs) so that they may be repeatedly used in the database. In other words, this set would act as a sort of semantic alphabet or defining vocabulary.

The intuitive criteria for assigning descriptors to phraseological units based upon the most common lexemes may seem quite simplifying and impoverishing, since the most frequent lexemes do not necessarily comprise the most common categories. Some researchers warn of such problems as polysemy of lexemes, which are newly acquired meanings used in defining vocabulary of their combinations or collocations. Additionally, defining vocabulary should also include some abstract concepts such as *property*, *phenomenon*, *quality* which are not among the most usable words (Xu, 2012). For the experiment's sake and for reasons of simplicity, we constrained ourselves at this stage to the most common ones. Polemic as it might seem, we decided to impose this restriction for another two reasons: 1) the descriptors we use are not definitions *sensu strictu*, but a mere list of semantic features, free of figurative meanings and thus with a minimum of polysemy; and 2) an experiment involving two or more level synoptic schemes based on the most frequent descriptors has not been conducted yet, and thus seems worth an attempt.

3.1 *One-level synoptic scheme*

For one-level classification, a straightforward frequency criterion was used: the 50 most frequent descriptors were considered candidates for the most general categories: *absurdity*, *accuracy*, *annoyance*, *astonishment*, *commitment*, *conflict*, *courage*, *criticism*, *damage*, *danger*, *death*, *deceit*, *despair*, *destruction*, *difficulties*, *diligence*, *dishonesty*, *efforts*, *end*, *exhaustion*, *experience*, *failure*, *fear*, *fight*, *haste*, *incaution*, *inefficiency*, *injustice*, *involuntariness*, *lack*, *laziness*, *limitation*, *loss*, *luck*, *nonsense*, *obstacle*, *passivity*, *poverty*, *power*, *problem*, *rage*, *resistance*, *risk*, *secret*, *speed*, *stealth*, *surprise*, *trouble*, *uncertainty*, and *visibility*. To avoid arbitrariness, the number choice was based on the total amount of descriptors and is argued further.

Having compiled a list of the most common descriptors in this manner, we determined that several entries in the database did not contain any of them, which is why we had to reconsider the possibility of completing some entries with one of the pivot (most frequent) descriptors. This completion task unveiled some lacking important features in the PUs descriptions, and allowed us to reduce the subjectivity of the first-round descriptors attribution. This lack of correspondence in some of the languages can be explained by originally intuitive annotation, which might seem extremely subjective and incomplete. However, the more these entries are processed, the clearer the difference between *contours* and *noise* (i.e., between the frequently and scarcely used descriptors) becomes. Afterward, all the descriptions were revised according to the established pivot categories, in accordance with the mentioned Method of Lexicographic Portrait.

A simple analysis of the pivot descriptors list could help to unveil important features: some of the most frequent descriptors are quite synonymic (*absurdity* and *nonsense*, *stealth* and *deceit*), which induced us to make a further revision of the semantic description. On the other hand, using synonyms in a synoptic scheme makes some sense, since some users may want to search the needed phrase by the idea *absurdity*, and some of them, by *nonsense*. This dilemma is thus open to debate.

It is worth noting that some descriptors constitute antinomian pairs (like *risk-caution*), but most of them do not. For example, although the descriptors *poverty*, *death*, and *criticism* are among the most frequent ones, their antonyms *richness*, *life*, *praise* are not on the list. Thus, it seems reasonable to use in further dictionary structure antinomian conceptual pairs, which will allow us to increment the efficiency of hyper-textual links and simplify the synoptic scheme by merely reducing the number of concepts. Ideally, the number of descriptors at each level should be equal to the root of the nth degree, i.e.:

$$d = \sqrt[n]{N}$$

(where ***d*** is the number of descriptors at each level, ***n*** corresponds to the number of levels and ***N*** stands for the whole number of descriptors). We note that this formula is aspirational and approximate in nature, because some descriptors may refer to several groups simultaneously, (i.e., ***d*** in practice will result in a greater number than the one calculated in theory).

3.2 *Two-level synoptic scheme*

As explained above, while a one-level synoptic scheme classifies the PUs into simple groups, a two-level scheme is meant to subdivide them into groups and subgroups.

We should also clarify the reason why we have used 50 descriptors on the top level, and not ten, as Pollux did. At first glance, ten descriptors provide quite a scarce variety of very broad characteristics, unhelpful in searching specific meanings, whereas as many as 2,000 would be extremely detailed and difficult to keep in mind at a time. Having about two thousand descriptors, which is coincidentally similar to the conventional number for defining vocabularies (Xu, 2012), we calculated that a scheme with ten initial descriptors would contain sections with 200 subcategories, although 200 descriptors are still quite difficult to process. We could keep adjusting the categories and subcategories this way to get them more balanced. A scheme with 20 initial descriptors, for example, and 100 further subcategories, appears much more manageable. Ideally, the size of the first level group, as well as that of each subgroup, should be equal or at least similar. For example, in this case we use 50 descriptors at each level, 50x50 makes 2,500, which is an amount commensurable with descriptors' set.

The right mathematical tool for equilibrating the scale of each level is to base it on an amount that is equal to the square root of the number of total descriptors. For 2,000 descriptors, the square root would be 44.72. The order of the root corresponds to the number of levels in the synoptic scheme: if we needed to organise the synoptic scheme in three levels, we should extract the root of the third degree, which would be around 12.59, a number that is manageable to work with and to perceive simultaneously in human operative memory. Still, this suggestion is to be subject to a separate experiment, which will be our goal for the next research.

4. Results and discussion

Having implemented the described methods, a two-level dynamic synoptic scheme for the dictionary has been compiled. Whilst the first level of the synoptic scheme is being compiled in accordance with a straightforward strategy (selecting the most frequent descriptors), the second-order descriptors selection is to be based on relations amongst them, beyond the quantitative parameter. One possible way of establishing semantic relations among descriptors is by analysing the personal perception of the affinity degree or by using thesauri. Both methods seem to be quite subjective and time-consuming. A personal perception, as well as actual thesauri, are quite arbitrary from the point of view of objectivity. Thus, our challenge was to establish an objective criterion applicable to computational technique.

For this purpose, we decided to explore the hypothesis that the categories (pivot descriptors), and their subcategories are likely to be placed within the same entries. In other words, we assume that descriptors and sub-descriptors have regular intersections within the same entries. For this reason, if a pivot descriptor A and descriptor B are found within the same dictionary entry more than once, B is automatically considered as a subcategory of A. For example, according to this principle, a pivot descriptor *secret* in Table 1 is related to descriptors *accuracy* and *correctness* (rows 1 and 3 respectively). Thus, we consider them as potential subcategories of the category *secret*. In order to exclude randomly or erroneously related descriptors, only those that appeared within the same entry more than twice were considered as candidates for subcategories. The number of co-occurrences could also be increased to three, making the degree of affinity at once more precise but less detailed.

A curious (and paradoxical) conclusion is that the scheme proposed for classifying phraseological entries is not a tree-based hierarchy with descending branches, but rather an encircled network, in which two different concepts can serve as mutual hyperonyms. For instance, the group *accuracy* includes *information* among its sub-concepts and, *vice versa*, the group *information* contains *accuracy* as a sub-concept, as seen in Figure 1. Each category forms out a node with outgoing rays which connect them with other nodes or isolated concepts. An analogy from our life could be that of subordination or competence overlapping, where a man might be the head of the household, whilst his wife might be his chief at work in their respective department, both of them being subordinated to the head of an interdepartmental supervision group. Traditional onomasiological dictionaries mentioned above are based rather on a tree-based hierarchy.

Despite the fact that the two, three, or multi-level scheme may seem difficult to handle, we tend to agree with Kawalya and Schryver, who affirm regarding a combined onomasiological and semasiological dictionary *Alphaconceptual+* that "in an electronic environment there is also no need for alphabetical indexes where the thematic information is listed. This considerably reduces the time and stress involved in moving back and forth connecting the words in the index to the words in the main body of the dictionary" (2013: 186).

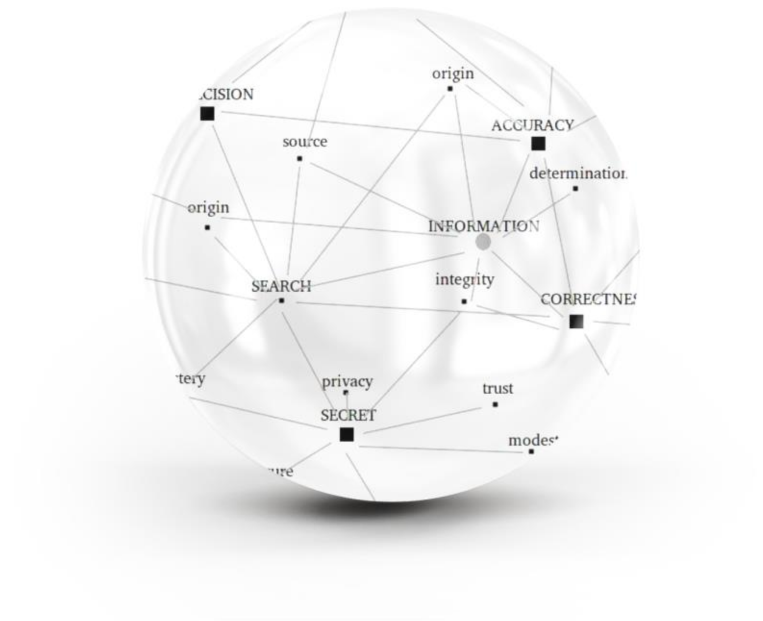
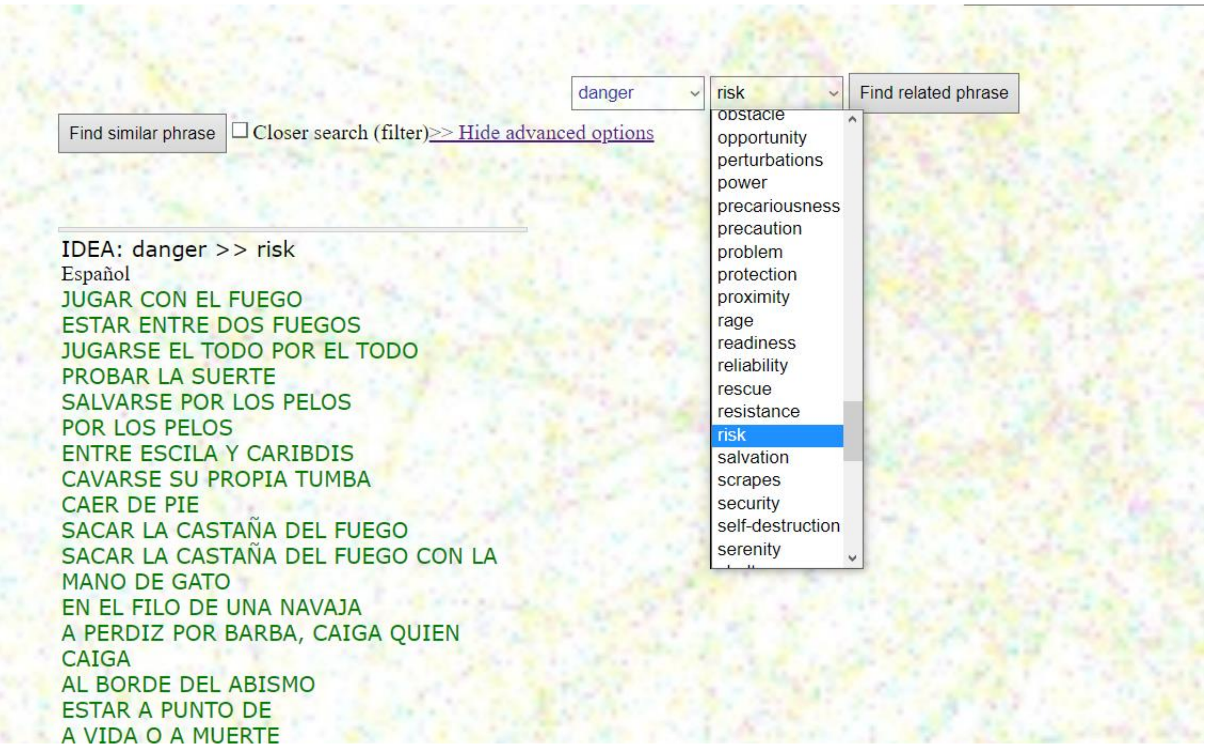


Figure 1: ***Illustration of sphere-formed hierarchy***

As a result of the experiment performed according to the proposed algorithm, a two-level synoptic scheme gets compiled "on the fly" during program performance with regard to the latest database update. For example, since the group *accuracy* regularly appears within the same entries with such descriptors as *adequacy, attention, awareness, belief, captiousness, clarity, cleanliness, cogency, compliance, confidence, copy, correctness, courage, courtesy, description, determination, disclosure, eye, honesty, head, information, integrity, knowledge, lack, match, meaning, news, oath, objectivity, obligation, overconfidence, perfection, precision, professionalism, prudence, punctuality, reality, reason, repeat, relevance, reliability, satisfaction, severity, similarity, sincerity, speech, straightness, timeliness, truth, uniqueness, wit,* and *word,* it is a valid candidate for subcategories. And, as the experiment shows, most of them indeed are.

Another useful property of descriptors' repetitions found in annotations is their possible usage for extracting semantically related PUs, generating phraseological thesauri, compiling lists of phraseological synonyms, or even translating from one language into another.

Thanks to the multi-label classification, a user query for expressing an idea by means of a phraseological unit may produce significant results taking into account the relatively modest dictionary volume (about 6.000 lemmas). For example, the query for the category *danger* along with its subcategory *risk* yields the 95 phrases in different languages, as in Figure 2, is shown, among which the following phrases in English: *to play with fire*, *in the crossfire*, *to go for broke*, *to chance one's arm*, *to have a narrow escape*, *by the skin of one's teeth*, *to dig one's own grave*, *on the brink of the abyss*, and *between a rock and a hard place*.



IDEA: danger >> risk
Español
JUGAR CON EL FUEGO
ESTAR ENTRE DOS FUEGOS
JUGARSE EL TODO POR EL TODO
PROBAR LA SUERTE
SALVARSE POR LOS PELOS
POR LOS PELOS
ENTRE ESCILA Y CARIBDIS
CAVARSE SU PROPIA TUMBA
CAER DE PIE
SACAR LA CASTAÑA DEL FUEGO
SACAR LA CASTAÑA DEL FUEGO CON LA MANO DE GATO
EN EL FILO DE UNA NAVAJA
A PERDIZ POR BARBA, CAIGA QUIEN CAIGA
AL BORDE DEL ABISMO
ESTAR A PUNTO DE
A VIDA O A MUERTE

Figure 2: ***IdeoPhrase output example***

5. CONCLUSION

The properties of vocabulary level can be analysed both by external data (such as texts, corpora) and internal data (descriptions in dictionary entries). Both of these datasets are potentially valuable for accomplishing classification tasks.

The experiment we conducted allowed us to apply an algorithm of an automatic compilation of 2-level synoptic classificatory scheme, as well as to put to the test its efficiency in an onomasiological, multilingual, phraseological dictionary performance. Semantic descriptions performed by using the Method of Lexicographic Portrait allows for illustrating the most common features and properties of each separate phraseological unit, and those of the phraseological dictionary in the whole. With respect to database preparation, the descriptors are assigned intuitively to each phraseological entry. Afterwards, automatic extraction of the most frequent descriptors performed in real-time (and *ad hoc* for a certain entries number) can be useful to extract many of the most common semantic features expressed by descriptors.

With the purpose of elaborating a two-level synoptic scheme, a criterion is needed for determining which descriptors should be considered categories, and which ones would qualify as subcategories. For 10,000 PUs and around 2,500 descriptors, we limited the categorical set to 50. Our algorithm consisted of the following observation: those descriptors that regularly appeared near a category within the same entries were candidates for subcategories. This algorithm allows for the creation of a handy synoptic scheme in real-time so that the list of automatically generated subcategories may include pertinent subsections. As a result, the synoptic scheme compiled on the fly in accordance with the algorithm described above yields a hierarchically based two-level classification where pivot descriptors may include each other as subcategories. The resulting structure underlying the synoptic scheme is a sphere-formed graph rather than a tree-based hierarchy. In paper dictionaries, a sphere-formed hierarchy would either be impossible to produce or would be full of overlaps.

Combining the dictionary of phraseological synonyms with an automatically generated concordance results is vital for the user to make the best contextual choice (connotational, pragmatic, and situational) for the given context of usage.

The analysis of the automatically extracted descriptors points to the need for revising and specifying semantic descriptions by eliminating synonyms among them and representing each idea as an antinomian pair.

A dynamic, synoptic scheme compiled on the fly might be valid for phraseological dictionaries, although this fact does not preclude the possibility that it should be applicable for common vocabulary, which is still to be proven in further experiments.

Acknowledgements

We would like to sincerely thank Iryna Shyyanova for having kindly provided her feedback and suggestions on the form and contents of the present study. We are also thankful to Maxym Ivanov for his critical review and constructive pieces of advice on program performance and interface design optimization.

References

Casares, J. (1994). *Diccionario ideológico de la lengua española*. Barcelona: Editorial Gustavo Gili, S.A.

Como se dice en. (2013). Traductor de palabras. Retrieved from https://www.comosediceen.com/

Cooper, M. C. (2008). Measuring the semantic distance between languages from a statistical analysis of bilingual dictionaries. *Journal of Quantitative Linguistics*, 15:1, 1-33, doi: 10.1080/09296170701794260

Darchuk, N. P. (2008). Kompiuterna linhvistyka (avtomatychne opratsiuvannya textu): pidruchnyk *[Computational linguistics (automatic text processing): manual]*. Kyiv: VPTS "Kyivskyi Universytet".

Darchuk, N. P. (2017). Mozhlyvosti semantychnoi rozmitky korpusu ukrainskoi movy (KUM). [Potentiality of semantic annotation in the Ukrainian Language Corpus]. *Naukovyi chasopys Natsionalnoho pedahohichnoho universytetu imeni M.P. Drahomanova [Scientific Journal of the M.P. Drahomanov National Pedagogic University]*, 15, 18-28.

de Schryver, G. M. (2003). Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography*, 16(2), 143-199. doi: 10.1093/ijl/16.2.143

Dicionário Priberam da Língua Portuguesa. (2019). Retrieved from https://dicionario.priberam.org/

Dobrovol'skij, D. & Baranov, A. (2011). *Semanticheskie otnoshenia vo frazeologii [Semantic relations in Phraseology]*. Dialogue.

Fazlyeva, Z. K. (2015). Types of interlanguage phraseological correspondences (based on English and Turkish languages). *Review of European Studies*, 7(9), 1-9. doi: 10.5539/res.v7n9p1

Fokin, S. (2019). Neural network pattern for enhancing functionality of electronic dictionaries. *Advanced Education*, 12, 150-158. doi: 10.20535/2410-8286.132940
IdeoPhrase. (2020). Multilingual Dictionary of Phraseological Synonyms. Retrieved from http://postup.zzz.com.ua/IdeoPhrase.html

IEDRA, Buscador de palabras. (n. d.). Retrieved from https://iedra.es

Kawalya, D. & de Schryver, G. -M. (2013). Introducing a new lexicographical model: AlphaConceptual+ (and how it could be applied to dictionaries for Luganda). *Lexikos*, 23, 172-200. doi: 10.5788/23-1-1210

Kenett, Y. N., Levi, E., Anaki, D. & Faust, M. (2017). The semantic distance task: Quantifying semantic distance with semantic network path length. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(9), 1470-1489. doi: 10.1037/xlm0000391

Kolomiiets, M. P. & Rehushevskyi, Ye. S. (1988). Slovnyk frazeolohichnykh synonimiv *[Dictionary of Phraseological Synonyms]*. Київ: Радянська школа.

Kunin, A. V. (1996). Kurs frazeologii sovremennogo angliyskogo yazyka: Uchebnik dla institutov i fakultetov inostrannykh yazykov*[A phraseology course of the contemporary English language: A manual for institutes and foreign languages faculties]*: Moscow: Vysshaya shkola.

Mellado Blanco, C. (2014). *La polisemia en las unidades fraseológicas: génesis y tipología*. Retrieved from:
https://www.academia.edu/8761132/La_polisemia_en_las_unidades_fraseológicas_génesis_y_tipología

OneLook Reverse Dictionary. (n.d.). Retrieved from https://www.onelook.com/reverse-dictionary.shtml

Pamies Bertrán, A., Iñesta Mena, Bálmacz, E. M. & Káloustova, O. (1998a). *Multilingual Electronic Phraseological Dictionary* "*AUTOFRAS*". Tempus Language Toolbox. Granada (CD-version).

Pamies, A, Balmacz, M. & Iñesta Mena, E. M. (1998b). Criterios para una fraseología onomasiológica automatizada. En A. Pamies Bertrán, & J. D. Luque Durán (Eds.), *Léxico y fraseología* (pp. 207-217). Granada: Mateo Ediciones.

Pawar, A. & Mago, V. (2018). Calculating the similarity between words and sentences using a lexical database and corpus statistics. *arXiv preprint arXiv:1802.05667*, 1-14. doi: 10.48550/arXiv.1802.05667

Piñero, I. (2011). Synonymy and antonymy in the framework of a dictionary of idioms/La sinonimia y la antonimia en el marco de un diccionario de locuciones. *LinRed. Lingüística en la Red, IX*, 1-27.

Popović, S. (2020). Onomasiological dictionary in bilingual phraseology. In J. Szerszunowicz & E. Gorlewska (Eds.), *Applied Linguistics Perspectives on Reproducible Multiword Units: Foreign Language Teaching and Lexicography* (pp. 141-149). Bialystok: University of Bialystok Publishing House.

Refranero Miltilingüe. (1997-2021). Retrieved from
https://cvc.cervantes.es/lengua/refranero/Busqueda.aspx

Reyes-Magaña, J., Bel-Enguix, G., Sierra, G. & Gómez-Adorno, H. (2019). Designing an electronic reverse dictionary based on two-word association norms of English language. In I Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubiček, S. Krek & C. Tiberius (Eds.). *Electronic Lexicography in the 21st Century: Proceedings of the eLex 2019 conference*, (pp. 865-880). Sintra, Portugal.

Rodríguez-Piñero, A. (2012). Variación y sinonimia en las locuciones. *Revista de Lingüística y Lenguas Aplicadas (RLLA)*, 7, 225-238. doi: 10.4995/rlyla.2012.1138

Roget, P. M. (2011). *Thesaurus of English Words and Phrases*. The Project Gutenberg: MICRA, Inc. Retrieved from http://www.gutenberg.org/cache/epub/22/pg22-images.html

Sierra, G. (2008). Natural language searching in onomasiological dictionaries. In M. Zock & C. Huang (Eds.) *Proceedings of the workshop on Cognitive Aspects of the Lexicon* (COGALEX '08) (pp. 32-38). Stroudsburg, PA, USA: Association for Computational Linguistics.

Skorokhodko, Ye. F. (1970). Linhvistychni osnovy avtomatyzatsii informatsiynoho poshuku *[Linguistic bases of information search automation]*. Kyiv: Vyshcha shkola.

Slovnyk Frazeolohichnykh synonimiv [Dictionary of Phraseological Synonyms]. (n. d.). Retrieved from http://rozum.org.ua/

SUM, Slovnyk ukrayinskoi movy. Akademichnyi Tlumachnyi slovnyk [Academic Explanatory Dictionary of the Ukrainian Language]. (2018). Retrieved from http://sum.in.ua/

Taljard, E. & Prinsloo, D. (2019). African language dictionaries for children — a neglected genre. *Lexikos*, 29, 199-223. doi: 10.5788/29-1-1518

Tsang V. & Stevenson S. (2004). Calculating semantic distance between word sense probability distributions. In H. Tou Ng & E. Riloff (Eds.), *Proceedings of the 8th Conference on Computational Natural Language Learning* CoNLL-2004 (pp 81-88). Boston, MA, USA: Association for Computational Linguistics.

Wordnet. (2021). *A Lexical Database of English*. Retrieved from
http://wordnetweb.princeton.edu/perl/webwn

Xu, H. (2012). A critique of the controlled defining vocabulary. *Lexikos*, 22, 376-381. doi: 10.5788/22-1-1013

Zhao, Z. (2018). Automatic classification of English vocabulary based on association rules. *International Conference on Intelligent Transportation*, *Big Data & Smart City* (ICITBS), Xiamen. 266-270. doi: 10.1109/ICITBS.2018.00075.