DOI 10.20535/2617-5339.2019.4.182393

UDC 811.111’33

Serhii Fokin
PhD in Translation Studies,
Associate Professor
Kyiv Taras Shevchenko National University,
Kyiv, Ukraine
ORCID ID 0000-0003-3920-1785
sergiyborysovych@ukr.net

QUERIES STRUCTURING FOR SOLVING GRAMMAR AND LEXICAL SEMANTIC PROBLEMS BY MEANS OF CORPUS TOOLS

Abstract. In spite of the rapid development of textual corpora along with that of the tools of processing them, many potential users are not fully aware of their utility for solving a wide range of text formulating problems. Beyond a quite straightforward strategy such as usage of asterisks and checking out collocations, the modern corpus tools are characterised by a high potential in solving also a wide range of semantic issues regarding grammar and vocabulary. Knowing the usage of search masks, part-of-speech, morphological and semantic tags are of great help in formulating pertinent queries. Although the semantic tagging in actual corpora is quite rare, it is a very promising feature; its application is still hindered by polysemy of semantic tags. Before being “translated” into a formal query language, a logical solution should be found on the basis of formal properties of linguistic signs by applying analysis of distributional (colligational and collocational) potentiality, substitution, calque, and morphological analysis. Substitution allows to extrapolate properties from one unit to another within the same semantic group; distribution offers the possibility to unveil several semantic components in the context, and, *vice versa*, to find out an expected lexeme by its hypothetical surrounding; calque is a powerful tool within the trial and error strategy for finding potential equivalents; analysis of frequency is helpful at the stage of results’ interpretation and evaluation of their reliability. Combination of these methods allows users to solve orthographic, punctuation, morphological, syntactic and lexical problems arising either in monolingual communication, translation or perform data mining.

Keywords: textual corpora; semantic analysis; distribution; search mask; query language; data mining.

1 INTRODUCTION

The development of computer tools is an undoubted blessing; however, it requires considerable efforts to handle them to the full extent. Since corpora may be considered as artificial intelligence tools, one of the challenges for human beings is to be able to communicate with this intelligence properly, and, in order to manage it, it is indispensable to learn to “translate” human ideas and questions into formalized language of queries so that the technique can interpret them correctly. Thus, comprehending new tendencies of development of these tools is fundamental for practical and theoretical, industrial and scientific goals.

Textual corpora have been recently becoming a more and more powerful tool for different philological purposes. Contemporary learners are already aware of the importance of checking out usage and correctness of collocates in a foreign language corpus and know the advantages it offers in comparison with traditional search engines like Google. Authors (journalists, writers, scholars) can benefit from corpora in order to clear up usage doubts or finding out stylistically marked phrases; teachers can find examples for elaborating didactic work; and, finally, the researchers are who may really take the best advantages of them. Actual corpora are becoming more and more specific, detailed and extensive, and understanding their structure, properties and possibilities requires knowledge on grammar, semantics, stylistics, text typology and other philological domains. While formulating of queries implies using a formalised language of the working environment, regular expressions and search masks. While trivial straightforward queries can help to check out the frequency of a word or collocation, a more complicated syntax of queries, by contrast, yields much more specific results and provides wider possibilities than just finding matching strings or using an asterisk instead of word ending. The new search tools can be applied not only for purely

scientific purposes, but also for solving grammar, lexical, stylistic issues, as well as performing translation from one language into another.

Due to an extremely wide potential of corpus tools, impossible to deal within an article, in the present paper we concentrate on approaches to solving semantic problems concerning grammar, lexical and punctuation levels on the stage of wording, editing, as well as revising written texts, which may occur both in translation process and in elaborating monolingual texts. The formalised corpus tools might seem superficial and lacking of relation with real content, but, as a matter of fact, contemporary computational resources are provided with a vast range of specific means to process the semantic level of language units.

Our objective, thus, is to summarise the types of problems regarding semantics which can be treated by means of textual corpora. Since the majority of corpora are monolingual, we illustrate the raised issues on a monolingual corpus, mainly on British National Corpus (BNC, 2004) and Hansard corpus of British Parliament speeches (Hansard corpus, 2016).

Multiple aspects on the question posed above have been treated recently. S.Sharoff focuses on uses of comparable corpora, particularly, with the purpose of choosing the right word for an expression (Sharoff, 2006, p. 28). G. Corpas Pastor suggests that an *ad hoc* compiled corpus may be of great value for testing variants of translation, particularly, cognats, along with their context (Corpas Pastor, 2004, p. 224–246). M. Chantal Pérez Hernández describes an algorithm for extracting terms based on syntactic structure of the sentence (Chantal Pérez Hernández, 2002). Hassani mentions such possibility of corpus as searching for collocates, analysing frequency, and even carrying out semantic search based on distribution (Hassani, 2011, p. 359). Although there has been a bunch of researches related to text mining, they are most commonly concentrated on scientific objectives of research projects, being practical recommendations comparatively infrequent. As T. Ah-Hwee observed in 1999, "Current text mining products and applications are still tools designed for trained knowledge specialists. Future text mining tools, as part of the knowledge management systems, should be readily usable by technical users as well as management executives" (Ah-Hwee, 1999, p. 5). This statement remains current nowadays.

2 METHODS

Since corpora are mostly used for attesting variants from the standpoint of their usage and correspondence to the norms, in this particular study we generalise the kinds of problem which go a step beyond the issues of preferences and formal correctness. In order to illustrate the manner in which corpora may be used in solving semantic issues, we decided to analyse questions posed in English Stackexchange Forum, one of the most visited forums with regards to the doubts and questions concerning different aspects and levels of the English language.

The methods chosen for answering questions comprise the usage of wildcard characters, part of speech and semantic tags, as well as other conventional signs used in BNC and British Parliament Hansard Corpus. The meaning of linguistic units can be established on the basis of distribution or colligation patterns and frequency analysis. We also propose in some cases to use substitution, calque and morphological analysis as a part of trial and error strategy for solving semantic problems. Analysis of frequency is helpful on the stage of results' interpretation. Nonetheless, before applying the mentioned methods, focused mainly on the formal features, a logical solution, based on experience and intuition, should be found. Some strategies can be generalised and formulated as algorithms or didactic recommendations.

3 RESULTS

Nowadays English Stackexchange Forum contains above 105,000 questions. Having analysed 500 of them, we have come to the conclusion that, at least, 100 could have been answered by using corpora. Taking into account that some of the questions appear to be out-of-topic or out-of-rules due to their incorrectness, the percentage of the corpus efficiency could even grow larger. It is natural that users of the forum prefer to get answers from human contributors, but, on the other hand, corpora possess a set of advantages: electronic corpora

obviate the problem since one can use them without the embarrassment of appearing ignorant. Second, tight budgets mean that sometimes it does not make economic sense to hire subject field experts. However, one can use corpora free of charge as much and as long as they wish. Third, since experts are human beings, they are not without their own limitations. They may forget something, put it wrongly, or worst of all, express their own views in a prejudiced manner. However, since corpora contain a collection of texts written by different subject field experts, they represent a far broader cross section of the expert views and, therefore, the difficulty can be cleared up (Hassani, 2011, p. 358).

Moreover, the questions which imply the cause of this or that phenomenon, as well as the explanation of their essence and links among them are hardly resolvable by means of corpora, since the data which corpora can provide are mostly general: "yes-no" answers (due to binary nature of the computational data structure), statistic data, as well as illustrative examples.

The reason we have chosen a resource of the English language is that of making it extensive for a wide philological community. The methods applied below are equally valid for other languages' corpora with similar characteristics including part of speech and semantic tags, possibility of using wildcards and regular expressions.

As a result, we have observed that semantic problems concerning language correctness and stylistically preferred usage can be solved in the following cases:

- when we know the searched word or phrase partially (the beginning of the word or its stem);
- when we know the morphological or syntactic mask of the searched word or phrase;
- when we are sure about highly probable immediate context of the searched word or phrase, particularly:
    - a specific lexeme;
    - a specific wordform of a certain lexeme (type);
    - a specific grammar phenomenon;
    - a specific punctuation mark;

Usage of these features implies specialised knowledge of linguistic levels and their properties, of the query language as well as a developed intuition and heuristic algorithm usage.

4 DISCUSSION

Let us analyse some of the typical issues which can be solved by means of corpus tools according to respective level they concern.

4.1 Orthography. Appropriate spelling seems to be the most formal aspect of a written text which can be easily revised by spell-checking software or smoothly solved by means of dictionaries. Nevertheless, some of the rules are quite variable and flexible, and English Stack Exchange Forum contains at present 1,134 questions classified as "Orthography". Some of them might bear relation to the word meaning. For example:

QUESTION: *"I'm trying to find the word, which I believe is something like "Vitral", like when someone is saying bad and/or inaccurate things, such as: 'This guy has been spewing a bunch of inaccurate vitral'"*(*English Language, & Usage Stack Exchange,* 2019).

SOLUTION: Let us assume that user is sure of the beginning of the word stem, and so are we. Consequently, we can perform a query with a wildcard symbol like "vitr*", which produces the following results: "vitro" – 531 occurrences, "vitriolic" – 71 occurrences, "vitrinite" – 60 occurrences, "vitreous" – 34 occurrences, "vitriol" – 24 occurrences (*BNC,* 2004). Then, it is not problematic to establish the truth by looking up definitions in an explanatory dictionary. What is more, if we scrutinize the results, we can immediately discard the second and the third for having suffixes of adjectives, since the searched lexeme is a noun. Being aware of this fact, we can make the query even more restricted with a POS-mask. More specifically, as a noun can be preceded by an adjective, the query we use might be as follows: "ADJ virt*". As a result, among the matching strings, the fifth one turns out to be the word we've been looking for.

4.2 Punctuation rules are often semantically dependant. Some of the rules are quite formal, but in most of the cases it is necessary to take into account the type of clause, the syntactic role, that is, grammar semantic analysis is needed. Let us consider an example:

QUESTION: "*Is my comma usage correct in this sentence? 'When there is too much carbon in the atmosphere, too much heat is trapped from the Sun's light rays, dramatically increasing the global temperature'*" (*English Language, & Usage Stack Exchange,* 2019).

SOLUTION: While most grammar guides recognise that subordinate adverbial clauses which precede the principal ones should be separated by comma (Oxford Guide to English, 2002), the adverbial phrases with gerund are not specified among the rules. One can use a more detailed guide of English or try to establish the truth by querying the corpus. More concretely, the query ", ADV * ing," that is, the mask which comprises the following components: comma + adverb + word ending in -ing, outputs 2,100 results, while the same query without comma yields a ten times greater amount: 21,216. However, this does not necessarily mean that all the results correspond to adverbial clauses with gerund. Consequently, we need to make a deeper query: the latter contains 126 occurrences of the bi-gram "desperately trying". But in which of them this bigram is adverbial phrase with gerund and not a part of a continuous tense? We need to find a way to discard the latter from the results. This fact can be established thanks to the mask "_vb* desperately trying", where "_vb" stands for forms of the verb "to be". In response to this query BNC outputs 45 results. Although, this mask does not unveil the cases of the verb "to be" separated from the structure by other possible words (such as adverbs or adverbial phrases), this lack can be offset by using asterisks instead of hypothetical intercalated words. Finally, the query "_vb* * desperately trying" gives another 7 results. Cases of separation by two or more words are much fewer, so we can perfectly neglect them. Thus, gerund does not indicate a parallel action, being a part of a continuous tense, in, at least, 52 instances. Having discarded these and other irrelevant results, we observe that in 37 of them the -ing form designates a parallel action to another one, expressed in the predicate of the principal clause. For example:

*She tried ineffectually to push him away, **desperately trying** to remember whether or not she had spoken those fateful words aloud. If he realised the effect he had had on her -- was still having on her...* (*BNC,* 2004, desperately trying, 126).

Once discarded the more noticeable examples of complex object and some occurrences of the verb "to be" separated from the participle by other words, we are fully aware that the usage of comma to separate adverbial phrases with gerund is preferred, still, not without exceptions:

*There was this idiot, sailing along **desperately trying** to simulate an atmosphere of... So you want to be an actor?* (BNC, 2004, desperately trying, 126)

4.3 Grammar rules. Some of the grammar rules are not strict and exhaustive. For example, according to "Oxford Guide to English Grammar", some of the nouns in plural are followed by verbs in singular when they denote games (*Oxford Guide to English,* 2002). However, that is not necessarily true for all the instances. Thus, it is not unreasonable to check out some cases in corpus: the aforementioned Guide enumerates among them "billiards", "darts", "dominoes", "draughts", but not "checkers". Since the mentioned term denotes a type of game, it might perfectly fit into the mentioned rule, although it is not mentioned in the rule and still has the form of plural. Indeed, the noun in question turns out to be quite polemic, as the queries "checkers are" and "chequers is" give quantitatively the same results: five instances of each one. Therefore, to clear out this doubt, we need to use more complicated masks of search involving morphological descriptors, based on colligation pattern: a verb is often preceded by a noun agreeing with it in gender and number. Particularly, the query "checkers _v?z*", that is, "checkers" followed by a verb in third person singular, yields only 8 results, while for "checkers verb" we obtain 77 results.

4.3.1 Morphological issues. By using specific query language one can clear out his or her doubts which go beyond classical grammar rules. Let us observe several examples.

QUESTION: *Can one use "man" like one can use "woman" as an adjective?* (*English Language, & Usage Stack Exchange,* 2019*).*

SOLUTION: To answer this question, we should query above all for fragments whose morphological mask is "man + NOUN", that is to say, where the noun "man" is used before another noun, according to a typical colligation pattern. A noun which precedes another noun is more likely to have properties of an adjective being an attribute. This particular query produces 1,031 results, among which the most frequent are the following:

"man hours" 71 occurrences
"man city" 61 occurrences
"man days" 13 occurrences
"man right" 13 occurrences
"man show" 11 occurrences (BNC, 2004)

Logically, our answer to the question posed above is definitely positive.

QUESTION: *Recently on IRC I said this: 'I do not believe in proving the correctness of already constructed programs. I believe in formally deriving programs so that they be correct'. And I got almost instantly corrected: 'DijkstraGroupie: So that they are correct, you mean'. Checking on* Wikipedia*, I found the following: 'I want you to give this money to him so that he* ***have*** *enough for lunch (the conjunction "so that"' takes a subjunctive in formal English). What usage is correct, in this case?* (*English Language, & Usage Stack Exchange,* 2019).

SOLUTION: Having made a query "so that it", we find out that this combination occurs 1,643 times, while the same structure followed by verb in third person singular, that is "so that it _v?z*" produces only 662 results. That is to say, the usage of the mentioned combination with bare infinitive is preferred to that with personal forms.

4.3.2 Syntactic issues. One of the most typical syntactic usages of corpora is checking out the verb, adjective and noun government. Although the most classical structures do figure in dictionaries, such as "to apply for", "to be fond of", the dictionaries might not cover all the possible cases. The "Oxford English Dictionary", in spite of providing examples, does not recommend neither explains the prepositions along with semantic details depending on them which users might need when answering the following question.

QUESTION: *'The yearbook is made with love by Lisa, with contributions* ***by*** *Mary and Sal' or 'The yearbook is made with love by Lisa, with contributions* ***from*** *Mary and Sal'* (English Language & Usage Stack Exchange, 2019).

SOLUTION: According to BNC, "contribution by" occurs 98 times, while "contribution/s from", appears 319 times in corpus, being the latter, as it is logical, the answer to the question. An elementary statistical analysis indicates the preferred usage.

Contemporary corpora, however, have much more potentiality of resolving syntactic questions which may arise regarding even sentence structure. Let us see another example.

QUESTION: *Can a sentence have no verb except in what would otherwise be its noun phrase?* (*English Language, & Usage Stack Exchange,* 2019).

SOLUTION: Thanks to part of speech tags, one can try out different search masks of sentence structure and observe which of them are realisable in practice. For example, the query ". NOUN NUM NOUN .", separated by dots as indicated, gives 31 result of phrases which seem to be titles or announcements like "Number one remedy", "Congress I second", "DANGER 600 VOLTS", "PARTY BOOKINGS # OPEN 7 DAYS A WEEK TO PARTIES BY APPOINTMENT ONLY **", "Minimum 25 persons".**

Nevertheless, an example of elliptic sentences may also occur under the mask "NOUN ART NOUN": "Her voice was happy once more and Kate watched her run from the roo: Thanks, Dan. **Thanks a bundle**'.

Thereby, one of the conclusions is that non-verbal sentences are widely used in

announcement, titles and spoken speech. The last example comes to illustrate the phenomenon of parcellation, thanks to which a sentence from the punctuational point of view may be structured without a verb.

4.4 Lexical meaning. Analysis of lexical meaning is, perhaps, still the least used in corpora. However, the closest context may really yield useful details concerning the semantic meaning of a certain lexeme and, logically, be undoubtedly useful in data mining analysis. According to the law of semantic combination, two words constitute a correct combination or collocation if they possess a seme in common (Gak, 1998, p. 279). For example, to establish the distinctive features among the synonyms denoting "sunrise", "dawn", "daybreak" in Spanish, García Meseguer proposes a scheme based on distribution analysis (García Meseguer, 2006). In other words, the usage of typical distribution makes it possible to establish whether the lexeme in question is an event or a point of time (depending on the prepositions used before), whether it refers to a durative process or punctual event (for example, depending on adverbs such as "during" or similar). We can conclude that, although corpora do not provide users with exact definitions, they can clarify important semantic details. After all, as a famous Spanish philosopher Ortega observes, external links of phenomena (in its terms, "contorno") are likely to be more important to understand their essence than their internal features (Ortega, 1983, p. 490–491). Similarly, the next question about the meaning of a lexeme can be solved by analysis of its surrounding.

QUESTION: *Can "household" be treated as a person? According to Oxford dictionary* ***household*** *means: a house and its occupants regarded as a unit, e.g., 'the whole household was asleep'. The question in my mind is if it can be regarded as the agent of an action. For example, is the following sentence correct or not? This chart demonstrates the distribution of renting and owning accommodation by households in percentage during 1918 to 2011 in England and Wales. Thank you in advance* (*English Language, & Usage Stack Exchange,* 2019).

SOLUTION: Since agentive complement with preposition "by" can act as one of the possible markers of animate nouns, the query we can use is "by * household*" (asterisks stand for possible determinants and plural form). Indeed, we find 8 results in BNC, for example: "Although land-use decisions are made **by the household**, these may well not be made equally by all members of it (...)" (BNC, 2004, by * household*, 37).

Semantic analysis can be based on other markers, depending on the type of a specific problem. For example, Jensen mentions a list of such indicators as parenthesis, slash, "or", "also", "referred to as" (Jensen et al., 2012, p. 27).

An extremely helpful and promising resource for solving this kind of problem would be a semantically tagged corpus. Although these corpora do not appear to be abundant, one of them definitely deserves to be spotted out is the Hansard corpus of British Parliament speeches on the web-site of Brigham Young University (Hansard corpus, 2016). Thanks to SAMUELS project carried out between 2014 and 2016, the corpus is semantically tagged in a detailed fashion. The lexemes in the corpus are classified according to a 37 classes grid which, on their turn, are subdivided into subcategories and subsequently into smaller groups. Handling these tags can be particularly useful to answer questions as those that follow in the current item.

4.4.1 Finding out an appropriate lexeme. Although corpora tools are not yet likely to provide lemmas by definitions, as reverse dictionary do, for example, "One Look Reverse Dictionary" (One Look Reverse, n.d.), some of the words may be found by syntagmatic context. For example, among the questions asked in the English Stackexchange Forum there is one which could be easily solved in a corpus:

QUESTION: *What is the right way of saying that you put a bed-sheet on a mattress? Do you put it on, lay it on or spread it on a mattress? If it's neither of the 3 please tell me what is the right word. I'm confused m. Thanks!* (*English Language, & Usage Stack Exchange,* 2019).

SOLUTION: Preferred usage is, probably, the most common query employed by users of corpora, and a good part of questions is focused on searching an alternative or more precise lexeme

so that the sentence sound naturally and the lexeme be more pertinent. This procedure can be performed both by typing the potential expression in quotation marks on an Internet search engine and by testing the phrase in question in specialised corpora. By entering the phrase in question with an asterisk in lieu of the polemic word, it is possible to find the verbs convenient to the phrase and their respective frequencies. To make the query more specific and focused on the problem, we can take advantage of the part of speech tags in BNC by means of query: "VERB ART bedsheet".

Unfortunately, this query does not produce any result. What is more, the noun "bedsheet" itself occurs only 5 times in BNC, while the "bed-sheet" is limited to 8 hits. This problem does not appear to be a big deal if we resort to a trick which consists in substituting the rare lexeme for another one of the same semantic group. This method called "using similarity class" is also proposed by Sharoff (2006, p. 26). In the case of "bed-sheet", one of the semantically closest terms is "blanket". "Blanket" is not only a semantically related word, it is, precisely speaking, a co-hyponym of "bebsheet", as they refer to the same holonym: "set of bed-linen", which occurs 1,054 times in BNC (BNC, 2004, set of bed-linen, 1504). Then, the mask "VERB ART blanket" produces a much richer output, where, at least, 12 occurrences correspond to the verb "put", which results to be the most usable verb with the noun "blanket" in BNC, and, logically, with the indicated kind of nouns.

QUESTION: *What's the correct term for when a small problem becomes a big problem? An example would be what happens to Larry David on "Curb Your Enthusiasm"…* (*English Language, & Usage Stack Exchange,* 2019).

SOLUTION: It seems reasonable to start with queries like this: {AR:11}{AP:06:c} provided that tags indicated above mean respectively:

AR:11 Ideas: matter (n) 331430, idea (n) 235243, subject (n) 93182, conceive (v) 39090, concern (n) 37673, topic (n) 36512, occasion (n) 29331, affair (n) 26950, question (n) 20224, issue (n) 18338 (Hansard corpus, 2016).

AP:06:c Ideas: Increase in quantity/amount/degree: more (r) 1491834, also (r) 1208525, increase (v) 493064, increase (n) 427575, further (r) 322176, other (j) 218658, additional (j) 211399, extra (j) 151207, increased (j) 143274, increasing (j) (Hansard corpus, 2016).

However, the results obtained are not deemed to be highly satisfying, because among the first 100 items the only two hits matching the initial query contain the verb "increase". At the same time, the query corresponding to the notion of "problem", {AP:06:c}, produces only 21 results among which also the verb "escalate" is included.

The contributors of the "English Stack Exchange" forum suggest the following answers to the question: "escalate", "intensify", "get out of hands".

Thinking over the relative inefficiency of the query described above, we might think that an even better result would be achieved if we used a specific tag for the semantic of 'exaggeration'. The problem is that in the Hansard Corpus there are at least 5 categories which include the idea "exaggerate" ("the mind", "attention / judgement", "relative properties", "emotions/mood", "space"), and trying each one seems to be quite a time-consuming procedure. The complexity of using semantic tags does not discard their great potential in specifying queries.

As long as the question does not concern a specific term from a narrow technical domain, phraseological form may be useful as well, so, in this particular case we can use an onomasiological phraseological dictionary, which is, by its onomasiological nature, semantically tagged. For example, in the dictionary "IdeoPhrase" the tag "exaggeration" is associated with the following phraseological units: "to cost pretty penny", "too much noise about nothing", "to ask for the moon", "to make a mountain out of a hillmole" (IdeoPhrase, 2019). Thus, the phrases "to make a mountain out of a hillmole" and "too much noise about nothing" might match the initial query.

Whilst the semantically tagged corpora offer the possibility of searching lexemes by their tags, there might appears a need of inverting search, that is, searching of tags for a certain lexeme. If we had this possibility at our disposal, we could resolve the above mentioned doubt concerning the semantic meaning of household. Nevertheless, semantic tags are not a panacea, because they do not necessarily resolve the metonymic or metaphoric usage of a lexeme, and in the example like "land-

use decisions are made **by the household**" or similar the noun "household" would metonymically denote persons, but a great deal of metonymic usages are not registered in dictionaries, so they can hardly be used for semantic tagging.

4.4.2 Conceptual search in corpus. Though this issue rather concerns researches than practical usage, it can be used by journalists or other authors who try to find out the ways of representing, treating or describing a particular concept, notion or theme. For example, Tatsenko (2018) performs a successful research of grammatical parameters of the concept "EMPATHY" by analysing the noun and the verb "empathise" along with their grammar paradigms, adjectives "empathetic", "empathic", "empathising" in correspondence with respective slots ("agent", "patient", "instrument" etc.) (p. 148–152). Actual corpora are not provided with specific tags for different slots, that's why an analysis of the sense by researcher appears to be the only valid method. For those who are interested in deepen into semantic parameters of a concept, it seems promising to use wildcard symbols. For example, the query "empath*" can cover at once all the instances of a lexeme, including different parts of speech mentioned above. The semantic tags "AR:06:f" and "AR:13" in the Hansard corpus of British Parliament speeches matching the meaning of "empathy" (Hansard corpus, 2016) can allow sampling of instances where the concept is expressed by means of different lexemes (well by synonyms, either by metaphors or metonyms) or finding lexemes directly associated with the concept in question.

4.4.3 Translation from one language into another. Many doubts in respect to derivational model may be solved in monolingual or bilingual explanatory dictionaries. Though, their entries are sometimes scarce and insufficient, while some of them contain extremely detailed descriptions which do not allow making out the real usage. For instance, the Dictionary "Multitran" is exceptionally popular among the translators who work with many language pairs including Russian and Ukrainian. Its translations proposed for the Russian term «дезодорант» are, among others, the following: "deodorant", "antipersipiant", "BO juice", "deodorizer", "antibromic", "deo spray", "odor eliminator", "odor scavenger", "pit stop", "body deodorant" (*Multitran,* n. d.). In such untenable situation it does not surprise that questions arise like this: "Is it deodoriser/deodorizer/deodouriser/deodourizer? In British English as well as American" (*English Language, & Usage Stack Exchange,* 2019). In order to choose among them a really usable term, we make the query: "deodor*", which gives a total of 83 results. Among them, 78 occurrences correspond to "deodorant", which constitute the great majority. Similarly, if we translate "соковыжималка" from Russian into Spanish, we will find a set of variants in "Multitran": "expimedera", "exprimidero", "exprimidor", "exprimidora" and others (Multitran, n. d.). Since these term possess a stem in common, by introducing the query "exprimid*" in a corpus, such as *CREA* (*Corpus de Referencia del Español Actual,* 2008) we establish that "exprimidor" is predominantly frequent.

Although many new dictionaries and term banks have appeared recently and those already in long existence have in some cases been subject to important changes, they still tend to disappoint specialized translators who hope to find all data they require in lexicographical or terminographic resources.

Putting aside bilingual corpora, which turn out to be rare for the most part of language combination, we would like to illustrate how monolingual corpora can be used for searching translation both for separate lexemes and phrases. We can observe that translation professionals in their practice resort to calque and check them out in search engines. However, once performed in a representative corpus, this procedure turns out to be definitely more effective than looking up in dictionaries because of the great amount of neologisms, which are much more likely to be widely used in web-resources long before being included in dictionaries. Given these circumstances, there is nothing better than testing the variants in specialised corpora. The advantage of the corpora is not only that of a thoroughly selected and verifies texts, but also that of looking up by means of wildcard symbols and regular expressions.

Let us see another example concerning translation of the bibleism "Don't look a gift horse in the mouth" one can only type in most corpora "caball* regal*". The indicated search is somewhat more appropriate than using specific forms, because several variants often exist for many idioms, and concerning the above indicated phrase, the possible variants in the *CREA* are "A caballo regalado no se le miran los dientes", "A caballo regalado no se le mira los dientes", "A caballo regalado no le mires el diente", "A caballo regalado no le mires el dentado" (*Corpus de Referencia del Español Actual,* 2008). Seeing that it is difficult to achieve an accurate literal translation, it is obviously reasonable to combine the procedure of calque with wildcard characters, as indicated in the example. This method is workable if we know the exact translation of one of the components, while we are not sure about the translation of the expression in the whole.

More issues concerning translation by using corpora are considered in "Computational lexicography and translation" (Fokin, 2017, p. 59–65).

5 CONCLUSIONS AND SCOPE FOR FURTHER RESEARCH

The formalised tools of text processing may seem extremely formal, whereas their potential in semantic analysis can defy most users' expectations. Numerous grammar semantic issues and lexical semantic problems can be clarified thanks to corpora tools. While these sources are primarily used for the purpose of checking spelling and usable collocations, these resources are also applicable for solving grammar rules doubts, interpreting lexical meaning, finding out translations and sources of intertextual fragments as well as performing other text mining thanks to analysis of lexical and grammar semantics.

While the actual studies related to researching corpora tools are quite prolific, the guides of using corpora for practical purposes are relatively rare. Our objective has been to focus on learners' and teacher's needs. Many problems can be solved by using specific queries, starting from literal revision of a doubtful phrase, passing by using POS, semantic masks along with wildcard characters, ending with calque when translating. Alternatively, we can use calque with wildcard characters instead of some words or word endings, a POS mask combined with the semantic one. The aforementioned methods may be used in a mixed way as a part of trial and error strategy in a heuristic manner.

The users need to be familiarized with a specific query language, and possess solid linguistic knowledge. While part-of-speech classification and wildcard characters can be commonly used and understood by most elementary schoolchildren, the knowledge of semantic annotation presupposes a deeper philological background. Thus, didactic materials for writers, journalists, translators and even high school students including problems (from the simplest to the most difficult) with their solutions should be welcomed.

Since the semantic tagging is so far a quite new property of the corpora, many of which lack so far of this utility, and those which do not, might be difficult to use due to certain polysemy and ambiguity of semantic tags, a mere intuition may result of greater value than a thoroughly described rule. Particularly, a substitution of a queried word for another one of the same semantic group may yield richer and clearer results than an initial query. It seems a promising direction in further researches to collect and generalise the intuitive heuristic tactics and algorithms applied by advanced users in order to formulate strategies of search as well as elaborate more specific automatic searching techniques.

REFERENCES

Ah-Hwee, T. (1999). Text Mining: The state of the art and the challenges. In N. Zhong, L. Zhou (Eds.) *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases* (pp 65–70). Berlin, Heidelberg, New York; Barcelona, Budapest; Hong Kong; London; Milan; Paris; Singapore; Tokyo: Springer. Retrieved October 29, 2019 from http://www.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf

Belyayeva, L. N. (2011). Korpusnaya Lingvistika i perevod: potentsial i ogranichenia [Corpus Linguistics and Translation: potentiality and limitations]. In Zakharov V.P. (Ed.) *Trudy mezhdunarodnoy konferentsii "Korpusnaya lingvistika-2011"* [Worsk of Intrernational Conference "Corpus linguistics-2011"] (pp. 86-91). Saint-Petersburg: Edition of Philological Faculty. [in Russian]

Bowker, L. (2000). A Corpus-Based Approach to Evaluating Student Translations. *TheTranslator, 6*(2), 183–210.

Bowker, L. (2001). Towards a Methodology for a Corpus Based Approach to Translation Evaluation. *Meta*, 2001, *46*(2), 345–364. DOI: https://doi.org/10.7202/002135ar.

Burnard, L. (2005). *Metadata for corpus work*. In M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 30–46.). Oxford: Oxbow Books. Retrieved October 29, 2019 from http://users.ox.ac.uk/~lou/wip/metadata.html#HDR

Chantal Pérez Hernández, M. (2002). Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento [Exploration of textual digitalized corpora for building terminological databases grounded on knowledge]. *Estudios de Lingüística del Español*, 18. Retrieved October 29, 2019 from http://elies.rediris.es/elies18/ [in Spanish]

Corpas Pastor, G.. (2004). Localización de recursos y compilación de corpus via Internet: aplicación para la didáctica de la traducción médica especializada [Localizing resources and corpora compilation via Internet: application for didactics of medical specialized translation]. In C. Gonzalo García, V. García Yebra (Eds.). *Manual de documentación y terminología para la traducción especializada* (pp. 223–258). Madrid: Arco/Libros. [in Spanish]

Corpus de Referencia del Español Actual. (2008). Corpus de Referencia del Español Actual (CREA) [Reference Corpus of Actual Spanish]. Retrieved October 29, 2019 from http://www.rae.es/recursos/banco-de-datos/crea [in Spanish]

Davies, M. (2004). *British National Corpus* (from Oxford University Press). Retrieved October 29, 2019 from https://corpus.byu.edu/bnc/

English Language, & Usage Stack Exchange. (2019). Retrieved October 29, 2019 from https://ell.stackexchange.com/questions/178126/express-interest-in-toward-to-something

Fokin, S. B. (2012). Dystrybutyvnyj analiz pry ukladanni dvomovnykh perekladnykh slovnykiv (na prykladi ukrayins "ko-ispanskykh vidpovidnykiv polya "osvita") [Distribution analysis in compilation bilingual translational dictionaries, case of Spanish-Ukrainian correspondances in the field 'Education']. *Problemy semantyky, prahmatyky ta kohnityvnoyi linhvistyky* [Problems of semanitcs, pragmatics and cognitive linguistics], 21, 490–500. [in Ukrainian]

Fokin, S. B. (2017). *Kompiuterna leksykohrafia i pereklad* [Computational lexicography and translation]. Kyiv: Taras Shevchenko National University of Kyiv [in Ukrainian]

Gak, V. G. (1998). *Yazykovye preobrazovaniya* [On linguistic transformations]. Moscow: Shkola yazyka i russkoy kultury. [in Russian]

García Meseguer, A. (2006). Nombres temporales "alba", "amanecer", "madrugada" [Temporal names 'abla', 'amanecer', 'madrugada']. *Punto y coma: boletín de los trarudctores españoles de las instituciones de la Unión Europea*, 100, 27–28. Retrieved October 29, 2019 from http://ec.europa.eu/translation/spanish/magazine/documents/pyc_100_es.pdf [in Spanish]

Hansard corpus of British Parliament Speeches. (2016). Retrieved October 29, 2019 from https://www.hansard-corpus.org/

Hassani, G. A. (2011). Corpus-Based Evaluation Approach to Translation Improvement. *Meta. Revue des Traducteurs*, *56*(2), 351–373. DOI: https://doi.org/10.7202/1006181ar.

IdeoPhrase. Onomasiological Multilingual Dictionary of Phraseological Synonyms. (2019). Retrieved October 29, 2019from http://postup.zzz.com.ua/IdeoPhrase.html#

Jensen, V., Mousten B., & Laursen A. L. (2012). Electronic Corpora as Translation Tools: A Solution in Practice. *Communication and Language at Work-ICT Tools and Professional Language*, *1*(1), 21–33. Retrieved October 29, 2019 from https://pdfs.semanticscholar.org/07b8/5e09bb1aad0dc74d1cff618f4704183caa92.pdf

Multitran. (n. d.). Retrieved October 29, 2019 from https://www.multitran.com/

One Look Reverse Dictionary.. (n. d.). Retrieved October 29, 2019 from https://www.onelook.com/reverse-dictionary.shtml

Orozco-Jutorán, M. (2018). Efficient Search for Equivalents at Your Fingertips – The Specialized Translator's Dream. *Meta. Revue des Traducteurs*, *62*(1), 1–241. DOI: https://doi.org/10.7202/1040470ar.

Ortega y Gasset, J. (1983). Sobre el fascismo [About fascism]. In *Obras completas* (T. III, pp. 489–497). Madrid: Alianza editorial, Revista de occidente. [in Spanish]

Oxford Guide to English. (2002). Retrieved October 29, 2019 from https://www.uop.edu.jo/download/research/members/oxford_guide_to_english_grammar.pdf

Sharoff, S. (2006). Translation as problem solving: uses of comparable corpora. In E. Yuste (Ed.) *Proceeding of Third International Workshop on Language Resources for Translation Work, Research and Training at LREC* (pp. 24–28). Magazzini del Cotone ConferenceCentre, Genoa, Italy. Paris: ELRA / ELDA (European Language Resources Association, European Language Resources Distribution Association).

Tatsenko, N. (2018). Grammatical parameters of the notional modus of EMPATHY concept lexicalised in modern English discourse. *Advanced Education*, 9, 148–153. DOI: https://doi.org/10.2053

Сергій Фокін. **Структурування запитів для розв'язання граматичних та лексико-семантичних проблем за допомогою корпусних інструментів.** Попри стрімкий розвиток корпусів текстів та знарядь для їхнього опрацювання, багато потенційних користувачів не повною мірою усвідомлюють і використовують потенціал корпусів у розв'язанні широкого кола проблем на етапі формулювання текстів. Окрім можливості доволі елементарного застосування (використання символу-зірочки замість пропущених символів або простої перевірки сполучуваності словоформ),

сучасні корпуси характеризуються набором корисних функцій для розв'язання широкого кола і семантико-граматичних, і лексико-семантичних проблем. Вміння використовувати маски пошуку, частиномовну, морфологічну і семантичну розмітку може стати вартісною допомогою у формулювання запитів. На етапі перед формулюванням запиту необхідно знайти логічне розв'язання проблеми на основі формальних властивостей мовних знаків та аналізу дистрибуції (як колігації, так і колокації), субституції, кальки та морфологічного аналізу. Метод субституції дозволяє проводити екстраполяцію властивостей однієї одиниці на іншу одиницю з подібної семантичної групи; метод дистрибуції дає змогу виявляти окремі семантичні компоненти в контексті і, навпаки, віднаходити відповідну лексему за її оточенням; калька є цінним знаряддям стратегії спроб і помилок для пошуку потенційних еквівалентів у перекладі; аналіз частоти вживання корисний на етапі інтерпретації результатів та оцінювання їхньої достовірності. Поєднання зазначених методів дозволяє користувачеві розв'язувати орфографічні, пунктуаційні, морфологічні, синтаксичні і лексичні проблеми під час формулювання текстів як в одномовному спілкуванні, так і під час перекладу, а також здійснювати добування даних за допомогою корпусів.

Ключові слова: корпус текстів; семантичний аналіз; дистрибуція; маска пошуку; мова запитів; добування даних.

Received: October 31, 2019
Accepted: November 18, 2019