The European Proceedings of
Social & Behavioural Sciences
EpSBS

Future Academy ISSN: 2357-1330

http://dx.doi.org/10.15405/epsbs.2018.04.02.14

WUT2018
IX International Conference "Word, Utterance, Text: Cognitive, Pragmatic and Cultural Aspects"

SOME MISTAKES OF LINGUISTIC CORPUSUSERS

Gulnara Lutfullina (a)*
*Corresponding Author

(a) Professor, Kazan State power Engineering University, Kazan, Russia, gflutfullina@mail.ru, 8 917 8 95 70 66

Abstract

The linguistic corpus users must be aware of receiving erroneous data. The first error is related to the word frequency in the diachronic perspective. It is necessary to use special formula in order to correctly calculate the word's frequency. The second error is related to the grammatical search. It is difficult to set correct search parameters. The third mistake is connected with lexical homonyms. It is necessary to be cautious when you meet all lexical homonyms. The fourth mistake is related to semantic features combination search. If you "play" with semantic features search, you can get ambiguous results. The user of the linguistic corpus should not rely entirely on the search results. There is always some "noise" in search results caused sometimes by contexts shortages. It is necessary to evaluate the received data based on your language competence, to correctly select and set the search parameters. An improperly compiled search can lead to distortion of real results: their exaggeration or understatement.

Keywords: Linguistic corpus, search query, word entry, context shortage, unremoved homonymy.

http://dx.doi.org/10.15405/epsbs.2018.04.02.14
Corresponding Author: Gulnara Lutfullina,
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

1. Introduction

Describing the user's work in the linguistic corpus, many linguists use the Latin expression *caveat emptor,* which literally translates as "the buyer should be aware". What exactly should the linguistic corpus user guard against? (Baranov, 2003). The study material are examples of National Corpus of Russian Language, hereinafter NCRL (National Corpus of Russian Language, 2005).

2. Problem Statement

The relevance of this study is to analyze mistakes in using the linguistic corpus in order to prevent them in future.

3. Research Questions

The subject of this analysis is to explore the most frequent mistakes identified in using National Corpus of Russian Language. The object of the study are typical mistaken search parameters and obtained incorrect results in using the linguistic corpus.

4. Purpose of the Study

The purpose of this article is to consider the most frequent mistakes identified by the use of linguistic corpus.

5. Research Methods

The methods of using linguistic corpus are now developing. Each scientist independently develops his own methods to compile searches in accordance with his scientific interests.

6. Findings

6.1. The first error is related to the calculation of the word frequency in the diachronic perspective. It is necessary to correctly calculate the tokens' frequency. Let's calculate, for example, the tokens' frequency of the ancient Russian word *перст / finger*. Using the corpus search results, we can get the following data: 141 tokens before 1800and 1069 tokens after 1900. We understand it can't be really correct because an ancient word can't be used more frequent recently. It is necessary to calculate the relative frequency based on the absolute quantity using a certain formula. According to this formula, the relative frequency is equal to: the tokens quantity should be divided by words quantity in the text and multiplied by 1 000 000. There are 12 636 732 tokens available before 1800. We get the frequency of 10. There are 208 953 262 tokens available after 1900. We get the frequency of 0, 05.

Let's consider the frequency of the ancient Russian word *завет / testament*. Using the corpus search results, we can get the following data. There are 1 095 tokens available in 623 documents with words volume 225 402 521 after 1900. According to the above formula, we get the frequency 12. There are 748 tokens in 421 documents with words volume 116 306 013 before 1900. We get the frequency 6.

http://dx.doi.org/10.15405/epsbs.2018.04.02.14
Corresponding Author: Gulnara Lutfullina,
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

6.2. The second error is related to the grammatical search. D. Beiber and S. Conrad explored corpus materials in teaching of English grammar and proved their effectiveness. According to their opinion, there are three types of important results for teaching grammar: 1) information about the frequency; 2) fixing comparisons; 3) associations between grammatical structures and words (Beiber, Conrad, 2012). Let's search, for example, propositions containing predicates of Future tense form in combination with time interval circumstances containing preposition *до / before, until.* Using the main corpus of National Corpus of Russian Language, we get the following search result – one proposition. There is one example containing an expression*Ябудуписать* / *I will write* in combination with time interval circumstance including the preposition *до / before* [(1)]. Thetotalsearchvolumeis115,645documents, 23,803,881propositions, 283,431,966words.

(1)Ларка, это письмо я *буду писать до тех пор,* пока не получу ответа на первое [НКРЯ]/ Larka, *I'llwrite*thisletter*until*Igetananswertothefirstone(NCRL).

There are no examples with Future tense predicates in combination with time interval circumstance including preposition *к/ by(котъезду, к приезду/ the departure, by the arrival)*. There is one example with Future tense predicate*появится*/ *will appear* in combinationwithtimecircumstance *к празднованию/ to the celebration,* expressing the precedence meaning.

(2) *Кпразднованию* 850-летияМосквывбулочных*появится*батонсинтригующимназванием «Веторон» [НКРЯ] / *To celebration of* the 850th anniversary of Moscow, loafs with an intriguing title "Vetoron" *will appear* in the bakeries(NCRL).

We tried to specify parameters. We used the searches with three words: an auxiliary verb *will* + a verb + *a preposition with a noun,* as well as other combinations. It is impossible to believe that in Russian Future tense predicates are so rarely used with the preposition*до /to*. Only as a result of numerous tests we managed to correctly set the search parameters: the first word is a verb, an indicative mood, the future tense + the second word is a preposition*до / to*. We found 27,999 tokens. There are three examples with Future tense predicates combined with timeintervalcircumstances containing preposition *до / before, until.*

(3)Этот процесс *затянется до октября,* и уже ясно, что с чеченской стороны его подпишет избранный президент. Это сделает документ ещё более значимым[НКРЯ] / This process *will drag on until October*, and it is already clear that on behalf of Chechen side the document will be signed by the elected president. This will make the document even more significant (NCRL).

While you are working with the linguistic corpus you can not completely trust it. It is necessary to evaluate the received data based on your language competence, to correctly select and set the search parameters (Hunston, Francis, 2000).

The main task of the linguistic corpus is not only to contain texts, but to contain correctly annotated texts. For example, a morphologically annotated corpus includes a morphological indication of speech parts for all words, which allows the user to quickly find a word of a desired speech part. However, over time, the corpus volume increased significantly, which led to drop in search performance. The main problem was that different grammatical analyzes attributed to the same word began to mix due to morphological homonymy. So, for example, the word *берет*means *a hat* and is considered as an inanimate masculine noun. Thesameword *берет* means *takes*andis also considered as an indicative mood of the verb*totake*PresentSimpletense, 3-d person. Due to these difficulties it is impossible to obtain search results

http://dx.doi.org/10.15405/epsbs.2018.04.02.14
Corresponding Author: Gulnara Lutfullina,
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

for a demand "verb of the masculine gender". You will encounter the same difficulties while searching the word of plural number *книжки / books*. Unremovedhomonymy could bring the user a lot of unpleasant surprises. Even in the search for a prefix "*re -"* the corpus gives out the word *пе-ре (re) -ц* (means*a pepper)*, although it has nothing to do with this prefix.

The linguistic corpus can be imperfect, so the user needs to remember and understand its features and limitations. There can always be omissions in the corpus, because in any live language, as a result of language play or word creation; there are lexemes, word forms and new meanings, which are called occasionalisms.

In the oral speech subcorpus you can find many examples of occasional or innovative word forms that are used to attract the interlocutor'sattentionand make statement informal or ironic. In this case, the annotation is more transparent and clear so that the user can easily find the necessary word form. Therefore, when drawing up a subcorpus, it is necessary to take into account the potential users and to facilitate interaction with the information source.

The "buyer" of the corpus, taking advantage of the "alien" corpus, consciously takes risks. The user takes risks that the corpus annotation may be imperfect. The corpus creators do certain compromise between theoretical knowledge and the possibilities of computer realization.

Therefore, the user should, on the one hand, be cautious about the search results received, on the other hand, the user should not forget that "not everything that seems, at first glance, is a mistake of the corpus".

6.3. The third mistake is lexical homonyms. For example, it is necessary to be cautious when you meet all lexical and grammatical homonyms. You shouldn't forget that unintelligible homonymy can produce ambiguous results (Kutter, Kantner, 2012).

There is a searchexampleof a word*лук/ an onion*inoral speech corpus on NCRL. As a result of search we have received the following examples:

(4) Ну/ тамбылтакойбоевойбизнесмениз*ВеликихЛук* / ДмитрийМатвеев / которыйизбиралсядепутатомобластногособрания / онбольшевсехпообещалвложитьденег / онисталлидером [НКРЯ] / Well, there was a businessman from *Velikiye Luk* / Dmitry Matveyev / who was elected as a deputy of the regional assembly / he promised to invest more money / he became the leader(NCRL).

(5) Аэтосамое… *алук*почёминтереснотогда? [НКРЯ] / And this is the most interesting... how much is *the onion* then? (NCRL).

(6) Знаете/ чтотакоеарбалет? *Усиленный лук.* Стрела его бьёт далеко/ пробивает рыцарские доспехи [НКРЯ] / Doyouknowwhatacrossbowis? *Reinforced crossbow.* His arrow beats far / knight armor pierces (NCRL).

In the example (4) we see the proper name, the location name *ВеликиеЛуки* / *Velikiye Luk.* In next example (5) the word *лук*means *an onion.* The word *лук*means *a crossbow* in the last example (6). You have to decide what is the most convenient to you.

http://dx.doi.org/10.15405/epsbs.2018.04.02.14
Corresponding Author: Gulnara Lutfullina,
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

6.4. The fourth mistake is to search for a semantic features combination. Also, you need be careful and attentive with the corpus. If you "play" with semantic features search, you can get ambiguous results. For example, if you specify semantic feature*частителa / "human body parts"* and *отрицательныечерты / "negative" features*, you can get the following examples:

(7) Икогдаясалатготовилвотиз*брюшексёмги*/ ятожебралморковьоттуда [НКРЯ] / And when I cooked salad with *salmon stomach* / I also took carrots (NCRL).

(8) Что означает «*перст указующий*» - знаменитый жест Древней Руси? [НКРЯ] / What does "*finger pointing*" mean - the famous gesture of Ancient Russia? (NCRL).

In the example (7) we see the word *брюшко*means *thestomach.* The word *перстуказующий* means *finger pointing* in last example (8). You have to decide if there is a human body part with negative feature among these examples. The "negative meaning" of the words in these examples is questionable.

6.5. The linguistic corpus user should not rely entirely on the corpus search results because there will be "noise" or context shortage to some extent.

1. Example of "noise": There are a large number of texts in NCRL, in which homonymy has not been removed. Therefore, in the search results of one word form, we meet contexts with a homonymous token. For example, if you are looking for a verb of imperative mood *рой / dig***,** you will also get a noun *пчелиныйрой / swarm (for example, "bee swarm*") into search results.

2. Example of "shortage": if we perform a search for word forms or if we incorrectly create a search for lexical and grammatical forms, the results will not show you the possible word forms, but only a part of them. Let's try to find the use of the quotation *рукописинегорят/ "manuscripts do not burn."* The search gives us 51 examples of this quotation. A lexico-grammatical search with the distance between words from 1 to 4 gives us 60 citations in the texts. Therearesuchinterestinginterpretationsamongthem.

(9) *Рукописи*,конечно, *негорят*,новосстанавливатьихохактрудно[НКРЯ] /*Manuscripts do not burn*, of course, but it is difficult to restore them (NCRL).

(10) Я, какдурачок, сиделиплакалнадними, неверя, что*рукописиивпрямьнегорят*[НКРЯ] / I, like a fool, sat and cried over them, not believing that the *manuscripts really do not burn*(NCRL).

(11) *Настоящие рукописи не только не горят,* но и не публикуются! [НКРЯ] / *These manuscripts are not only burning*, but they are not being published either! (NCRL).

An example of competent corpus use in phonetic research is the works of A. Piperski and A. Kukhto **(**Piperski, Kukhto, 2016). In their article, the authors automatically analyze the selectedsubcorpusincluding poems by ten poets from National Corpus of Russian language. They performed their search for word forms with stress variability. The word forms quantity lies in the interval from 30 to 200 word forms for different speech parts. In the article, a quantitative measure is proposed for estimating the overall variability of stress, independent of the corpus size.

7. Conclusion

Thus, we find that an incorrectly compiled search query can lead to distortion of real results: their exaggeration or understatement. This is something that should be aware of the corpus user.

http://dx.doi.org/10.15405/epsbs.2018.04.02.14
Corresponding Author: Gulnara Lutfullina,
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

References

Baranov, A.N. (2003) *Corpus linguistics*. Moscow: Progress.

Beiber, D., Conrad, S. (2012, January 29). Corpus Linguistics and Grammar Teaching. Retrieved from http://longmanhomeusa.com/content/pl_biber_conrad_monograph5_lo.pdf.

Hunston S., Francis G. (2000). Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English. Amsterdam; Philadelphia: John Benjamins Publishing Company.

Kutter, A., Kantner, C. (2012).
Corpus-based Content Analysis: A Method for Investigating News Coverage for War and Intervention. Retrieved from http://www.uni-stuttgart.de/soz/ib/forschung/IRWorkingPapers/IROWP_Series_2012_1_Kutter_Kantner_Corpus-Based_Content_Analysis.pdf

National Corpus of Russian Language (NCRL) (2005) Retrieved from http://www.ruscorpora.ru

Piperski, A., Kukhto A. (2016, June19). Intra-speaker stress variation in Russian: A corpus-driven study of Russian poetry. Retrieved from http://www.dialog-21.ru/media/3419/piperskiachkukhtoav.pdf.