UDC 81'322'42:004
DOI https://doi.org/10.32838/2710-4656/2022.1-2/26

Stepanov V. V.
Sumy State University

CORPUS VS NON-CORPUS: MAIN DISTINCTIVE FEATURES

The article reveals key principles to construct a linguistic corpus as a simplified model of communication. It is focused on them to distinguish between a true corpus and a random set of resources that is not a corpus itself. To instruct researchers in compiling corpora properly, such principles (representativeness, electronic format, annotation, software processing) are explained by the author.

Representativeness as a corpus property to be a compact analogue for real communication is regarded as the most significant attribute of corpora. Thus, taken resources must correspond to all features of discourse (texts are collected by age, sex, occupation, country of native speakers' origin, language variant, etc.). A single typology of corpora is illustrated within the representativeness rule.

Electronic format means that texts must be converted into a machine-readable form. Otherwise, resources will not be processed during the research. Specific techniques of e-conversion are discussed: scanning, printing, editing. Mechanisms of text encoding are analyzed.

In terms of annotation, corpus resources are divided into logical sections and marked with signs for linguistic and extralinguistic phenomena. The former covers verbal (semantics, grammar, syntax) and non-verbal (e.g. intonation, gestures, handwriting) aspects of communication. The latter includes discourse metadata: text authors' names, references.

Further, corpus software processing is discussed. Corpus managers as programs for analyzing corpora are indicated. Among them, the AntConc application potential for working with corpus resources is revealed: concordance, collocation, word list and so on.

Finally, ideas of virtual corpora are considered in the article. In particular, how they correlate with conventional corpora and what advantages they have. Briefly, basic advice for researching virtual resources is given.

Key words: *corpus, representativeness, electronic format, annotation, software processing.*

Problem and analysis of publications. Today, most linguistic studies are concentrated on use of illustrating research materials – generally so-called "corpora". For the last decade, there has been a great amount of researches based on corpus resources: discourse [3; 12; 21; 27], cognitive linguistics [6; 15; 25; 29], linguodidactics [4; 5; 7], etc. Irrespective of such a rise in studies, it is necessary to keep certain requirements of corpus compiling. They must be followed to make a research as authentic as possible. Otherwise, the study will fail.

On the other hand, there is an increasing number of PhD in Philology students who produce their own corpora within theses. Therefore, young scholars should be instructed in generating corpus properly to secure correctness of their linguistic research results. That defines the **article relevance**.

The research object is corpus. **The topic** is description of main principles of compiling true linguistic corpora.

The purpose is providing a detailed manual for scholars to construct a real corpus for their own studies. That is reached via the following **tasks**:

1) to define the notion of corpus and its differences from a random text set;

2) to enumerate and explain four main distinctive features of true corpora (representativeness, electronic format, annotation, software processing) and how they are kept within corpus compiling;

3) to give an example of corpus construction for a certain research.

Research. Any linguistic study must be conducted on live resources, which is generally called "discourse" or "communication". Obviously, discourse itself is limitless, and no scholar can learn the full scope of speech phenomena. Therefore, compact authentic communication samples have to be taken for scholars to complete their studies. Such samples are named corpora.

What is corpus? Corpus is a set of speech fragments selected for a specific research objective (in contrast to archive whose resources are collected randomly) [2, p. 15, 48–49]. Simultaneously, deliberate text selection is insufficient among corpus compiling requirements. A full-fledged corpus is only that set of resources which is arranged within

four rules – representativeness, electronic format, annotation, software processing.

Analyzing theoretical sources on corpus linguistics [17; 18; 20] makes a detailed explanation of these features and their role in compiling corpora.

Representativeness is corpus ability to be a communication mini-model that is maximally equal to reality. Since humanity has been developing constantly, no corpus can cover the total discourse boundaries. Subsequently, texts must be selected in that way when the limited content reflects a true speech character. For example, a medical corpus should include only medical texts; a British English gender corpus must contain both male and female utterances of UK origin.

The representativeness balance is a difficult task. Depending on researcher's aims, there may be produced different corpora with peculiar representativeness levels. Thus, the representativeness rule makes it possible to classify corpora by certain criteria. As an example, we can observe the corpus typology by V. P. Zakharov and S. Yu. Bogdanova [30, p. 16–25]:

a) all-purpose, special corpora (research objective);

b) dynamic, static corpora (resource refreshment pace);

c) oral, written, audial-and-visual corpora (communication form);

d) colloquial, literary, scientific, official, journalistic corpora (speech style);

e) literary, folklore corpora (genre);

f) full-text, fragmentary corpora (content volume);

g) monolingual, multilingual corpora (amount of languages);

h) learning, translation corpora (specialty);

i) free, paid corpora (commerce).

The above-mentioned criteria are not complete. Corpora may be classified by other principles as well.

Nevertheless, scholars focus mostly on corpora as to the research objective criterion, particularly on all-purpose corpora. The reason: resources of all-purpose corpora integrate total authentic range of genres and styles within a certain language; therefore, they are a reliable empirical material to prove or disprove research hypotheses [26, p. 118–120]. Thus, for the highest representativeness, special corpora should be primarily arranged via all-purpose bases and secondarily supplemented by texts from other sources.

Apart from the rank of authentic resource bases, some other features have to be included by a researcher to provide the optimal representativeness of own corpora. In particular, L. Flowerdew [9] and A. Koester [13] offer to mind the following factors in selecting corpus texts:

a) research object (what the corpus will study, in which language variant, genre, style, discourse);

b) features of communication participants (age, sex, occupation, language competence, place of birth and residence, etc.).

In such a manner, texts are collected. They are recorded by the main ethical rule: the information has been collected with speakers' voluntary consent, and all personal data have been removed or changed (public texts are an exception) [11, p. 157–160; 18, p. 154–168]. Further, the sample is digitized.

Electronic format is an integral part of corpus methodology of language study (because only digital data may be analyzed by software tools). Texts are digitized via scanning, printing or editing [19, p. 62–63; 24, p. 14].

Scanning is the quickest but the least accurate way of electronic conversion: texts can be recognized with failures (because of handwriting, paper deterioration), which makes you correct mistakes. The same concerns editing: the downloaded texts may be badly pre-formatted in the Microsoft Word application. Therefore, printing is the longest but the most reliable way of text digitization.

The next step is text encoding. Its significance is huge: improperly encoded files will not be processed by software, and the study itself will fail. We advise text preparation in the Notepad program. All resources are saved in the *txt* format via the Unicode system (UTF-8) [23, p. 32–35].

Then, texts are annotated. **Annotation** is segmentation of texts with producing certain marks. The content is subdivided into sentences, paragraphs, dialogic blocks with signs for linguistic and extralinguistic phenomena [18, p. 29–35; 23, p. 35–36]. Linguistic marks describe text elements in verbal (semantics, grammar, syntax) and non-verbal (intonation, gesture, handwriting, etc.) aspects. Extralinguistic marks provide metadata: names of text authors, references. They are indicated in each text as headlines / endings or given in a separate corpus file.

Some scholars regard linguistic marks as optional. According to J. Sinclair [24, p. 21] and M. Nelson [19, p. 63], it is text representativeness that is prior in corpora: authorship rather than lexeme attribution is the most important. Subsequently, metadata as extralinguistic signs are obligatory in compiling corpora while linguistic ones may be omitted.

The encoded and annotated corpus is ready for use. The last step is **software choice** to work

with resources efficiently. Such an application is a corpus manager.

Corpus managers are programs where you can adjust parameters of searching for necessary units in the total corpus. They generate a concordance – a list of all contexts where the searched unit is detected [2, p. 42–44].

Today, there are many corpus managers. Among the most popular ones, the AntConc application [1] is highlighted. It was designed by L. Anthony for the Windows, Linux and Macintosh operating systems. L. Anthony's lectures [14] show advantages of using this software. Firstly, units can be searched by cases, flections or lexeme limits. Secondly, results are sorted by the left-side or right-side valence. Thirdly, the data are copied in a separate dialogue box to compare them with adjacent search results. Finally, they can be selected and sent to a text processor for further research needs.

Along with concordance, the AntConc has other tools, e.g. collocation and word list. As means of the quantitative analysis, the former determines the amount dominance of interlexeme combinations while the latter defines the word frequency hierarchy among all other similar units within the whole corpus. The results are analyzed with a conclusion what specific speech preferences are traced among language native speakers.

Moreover, corpus is interpreted not only as a text bank. The World Wide Web is also increasingly regarded as a global collective set of resources, which comprises the communicative humanity experience from the historical perspective [26, p. 124–125]. Thus, the Internet is often called the virtual corpus.

According to W. Fletcher, the virtual corpus dominates over all other corpora significantly [8, p. 25–45]. Its advantages:

a) no local limits (the Internet is available everywhere);

b) no extra costs for software (each computer has a simple browser);

c) once-a-second resource refreshment (in contrast to conventional corpora whose content is edited annually);

d) coverage of all humanity spheres (the virtual inventory includes texts of those genres and styles that may be still not indexed by other corpora);

e) presence of textual as well as audial-and-visual resources (while conventional corpora usually possess only texts).

Subsequently, it is the virtual corpus that is the most ideal representative model of discourse while other corpora only approximate to reality.

Taking into account the advantages, the Internet is considered as both corpus and base for producing new corpora. Its content is indexed by search engines (Google [10]) whose functionality is similar to that of corpus managers: requests generate concordances of web pages with a searched unit. The results are sifted by data language, their place and time of creation, file format and so on. The resources are downloaded, converted, annotated, encoded and used for research aims.

Apart from search engines, linguists use specially developed services to work with the virtual corpus. A. Lüdeling, S. Evert, M. Baroni [16] and A. Renouf., A. Kehoe, J. Banerjee [22] advise the WebCorp service [28]: it generates concordances and produces word lists within the Internet. However, priority is traced in search engines rather than such services. The reason: service results are post-processed search engine ones (therefore, the WebCorp cannot find resources if they are not indexed by Google [16, p. 17]). Simultaneously, the search engine functionality increases via adjusting parameters of the network firewall, which may raise request hits. Thus, it is search engines that are the best corpus managers to work with the virtual corpus.

Let us observe an example how all above-mentioned rules work in compiling a corpus. The situation: we would like to research the POLITICS concept actualization in the modern American English discourse of presidents.

1. Representativeness. In the Internet, we look for only those American English utterances where the word "politics" occurs as a name of the corresponding concept. They must concern only USA presidents and be as various as possible from many people (for America, the today's number of presidents is 46). The gender factor is omitted: only males were USA leaders.

2. Electronic format. All found utterances are downloaded and edited in the Notepad program. They are saved as *txt* files in the UTF-8 code.

3. Annotation. The utterances are subdivided into sentences and paragraphs. Additionally, extralinguistic marks are indicated: headlines of each file is accompanied with authors' names and references.

4. Software processing. The AntConc is used to generate concordances by the search request "politics". The hits are analyzed for further research aims.

Conclusion. Therefore, corpora are simplified authentic models of communication. They differ from a random text set by four principles: representativeness, electronic format, annotation,

software processing. These rules may be used to check whether samples are true corpora or to produce new corpora via existing utterances. It can be engaged in studies of scholars and students (PhD, Master's, Bachelor's degrees) to prove or disprove hypotheses in their papers.

References:

1. AntConc (corpus manager by L. Anthony). URL: https://www.laurenceanthony.net/software.html
2. Baker P., Hardie A., McEnery T. Glossary of corpus linguistics. Edinburgh : Edinburgh University Press, 2006. 187 p.
3. Baker P., Vessey R. A corpus-driven comparison of English and French Islamist extremist texts. *International Journal of Corpus Linguistics*. 2018. Volume 23. Issue 3. P. 255–278. URL: https://doi.org/10.1075/ijcl.17108.bak
4. Boulton A. Corpora in language teaching and learning. *Language Teaching*. 2017. Volume 50. Issue 4. P. 483–506. URL: https://doi.org/10.1017/S0261444817000167
5. Chen M., Flowerdew J. A critical review of research and practice in data-driven learning (DDL) in the academic writing classroom. *International Journal of Corpus Linguistics*. 2018. Volume 23. Issue 3. P. 335–369. URL: https://doi.org/10.1075/ijcl.16130.che
6. Divjak D., Arppe A. Extracting prototypes from exemplars. What can corpus data tell us about concept representation? *Cognitive Linguistics*. 2013. Volume 24. Issue 2. P. 221–274. URL: https://doi.org/10.1515/cog-2013-0008
7. Dupont M., Zufferey S. Methodological issues in the use of directional parallel corpora. *International Journal of Corpus Linguistics*. 2017. Volume 22. Issue 2. P. 270–297. URL: https://doi.org/10.1075/ijcl.22.2.05dup
8. Fletcher W. Concordancing the web: promise and problems, tools and techniques. *Corpus linguistics and the web*. New York : Rodopi, 2007. P. 25–45.
9. Flowerdew L. The argument for using English specialized corpora to understand academic and professional settings. *Discourse in the professions: perspectives from corpus linguistics*. Amsterdam : John Benjamins, 2004. P. 11–33.
10. Google (the virtual corpus search engine). URL: www.google.com
11. Hunston S. Collection strategies and design decisions. *Corpus linguistics: an international handbook (Volume 1)*. Berlin : Walter de Gruyter, 2008. P. 154–168.
12. Ji M. A quantitative semantic analysis of Chinese environmental media discourse. *Corpus Linguistics and Linguistic Theory*. 2018. Volume 14. Issue 2. P. 387–403. URL: https://doi.org/10.1515/cllt-2016-0040
13. Koester A. Building small specialised corpora. *The Routledge handbook of corpus linguistics*. London and New York: Routledge, 2010. P. 66–79.
14. Lectures by L. Anthony (use of the AntConc corpus manager). URL: https://drive.google.com/open?id=1syjmVL2bTHAhnhi87LeLrbPcGfsixSFV
15. Lederer J. Finding source domain triggers: how corpus methodologies aid in the analysis of conceptual metaphor. *International Journal of Corpus Linguistics*. 2016. Volume 21. Issue 4. P. 527–558. URL: https://doi.org/10.1075/ijcl.21.4.04led
16. Lüdeling A., Evert S., Baroni M. Using web data for linguistic purposes. *Corpus linguistics and the web*. New York: Rodopi, 2007. P. 7–24.
17. Lüdeling A., Kytö M. Corpus linguistics: an international handbook (Volume 1). Berlin : Walter de Gruyter, 2008. 776 p.
18. McEnery T., Hardie A. Corpus linguistics: method, theory, practice. Cambridge: Cambridge University Press, 2012. 294 p.
19. Nelson M. Building a written corpus: what are the basics? *The Routledge handbook of corpus linguistics*. London and New York : Routledge, 2010. P. 53–65.
20. O'Keeffe A., McCarthy M. The Routledge handbook of corpus linguistics. London and New York: Routledge, 2010. 682 p.
21. Popova Ye., Yemelyanova Ye., Prikhodko N. Grammatical and lexical constituent of pre-election discourse. *Middle-East Journal of Scientific Research*. 2014. Volume 19. Issue 1. P. 48–51.
22. Renouf A., Kehoe A., Banerjee J. WebCorp: an integrated system for web text search. *Corpus linguistics and the web*. New York : Rodopi, 2007. P. 47–67.
23. Reppen R. Building a corpus: what are the key considerations? *The Routledge handbook of corpus linguistics*. London and New York : Routledge, 2010. P. 31–37.
24. Sinclair J. Corpus, concordance, collocation. Oxford : Oxford University Press, 1991. 179 p.
25. Tatsenko N. V. Grammatical parameters of the notional modus of the EMPATHY concept lexicalized in modern English discourse. *Advanced Education*. 2018. Issue 9. P. 148–153. URL: https://doi.org/10.20535/2410-8286.107093

26. Teubert W., Čermáková A. Directions in corpus linguistics. *Lexicology and corpus linguistics: an introduction*. London and New York : Continuum, 2004. P. 113–165.

27. Vaez Dalili M., Vahid Dastjerdi H. A contrastive corpus-based analysis of the frequency of discourse markers in NE and NNE media discourse: implications for a "universal discourse competence". *Corpus Linguistics and Linguistic Theory*. 2013. Volume 9. Issue 1. P. 39–69. URL: https://doi.org/10.1515/cllt-2013-0010

28. WebCorp (the virtual corpus service). URL: http://www.webcorp.org.uk

29. Yehorova O., Prokopenko A., Popova O. The concept of European integration in the EU-Ukraine perspective: notional and interpretative aspects of language expression. *Online Journal Modelling New Europe*. 2019. Issue 29. P. 53–77. URL: https://doi.org/10.24193/OJMNE.2019.29.03

30. Zakharov V. P., Bogdanova S. Yu. Corpus linguistics. Saint Petersburg: Saint Petersburg State University, 2013. 148 p.

Степанов В. В. КОРПУС VS НЕКОРПУС: ГОЛОВНІ ВІДМІННОСТІ

Стаття розкриває основні принципи конструювання лінгвістичного корпусу як спрощеної моделі комунікації. Наголошується, що витримка таких правил є корінним критерієм для розмежування істинного корпусу та випадкового набору текстів. Для правильного відбору ресурсів і укладання корпусів автор рекомендує дослідникам завжди дотримуватися чотирьох правил – репрезентативності, електронного формату, анотування, програмної підтримки.

Репрезентативність, тобто здатність корпусу бути максимально наближеною до реальності мініатюрною моделлю комунікації, визнається головним атрибутом корпусу. Так, дібрані ресурси мають відповідати всім параметрам спілкування (тексти вилучаються з урахуванням віку, статі, роду занятості, країни походження мовців, варіанту мови тощо). В рамках правила репрезентативності пропонується єдина типологія корпусів.

За електронним форматом, тексти повинні бути конвертовані в машинозчитувальну форму (інакше ресурси не будуть оброблені під час дослідження). Оглядаються конкретні прийоми електронної конвертації: сканування, друк, редагування. Аналізуються механізми кодування текстів.

З точки зору анотування, корпусні тексти членуються на логічні секції та маркуються на предмет прояву лінгвістичних і екстралінгвістичних явищ. Лінгвістична розмітка охоплює опис вербальних (семантика, граматика, синтаксис) та невербальних (інтонація, жести, почерк) аспектів комунікації. Екстралінгвістична розмітка зазначає дискурсивні метадані: автор тексту, посилання.

Програмна підтримка передбачає використання корпусних менеджерів для опрацювання корпусів. Зокрема, демонструється потенціал корпусного менеджера AntConc у виконанні досліджень: інструменти конкордансу, колокацій, вокабулярного переліку тощо.

Зрештою, розглядається ідея віртуальних корпусів: чим вони відрізняються від звичайних корпусів та які мають переваги. Стисло надаються ключові поради для роботи з віртуальними ресурсами.

Ключові слова: *корпус, репрезентативність, електронний формат, анотування, програмна підтримка.*