Teacher-student cooperation boosting grammar in the age of Big Data

Ekaterina Evgenievna Golubkova, *Marina Alekseevna* Salkova[*], *Olga Arkadievna* Machina[†], *Olga Nikolaevna* Pavlyuk, and *Aleksandr Vladimirovich* Trubochkin

Moscow State Linguistic University, English Language Department, Moscow, Russia

Abstract. The article focuses on the new language teaching and learning strategies offered by the use of corpora-based method. It gives an insight into how a corpus-based research incorporated into English grammar classroom activities not only clarifies the debatable aspects of traditional grammar rules but challenges students to launch their own language investigation, enlarging their professional scope and contributing to the development of their general and linguistic competences. The background for the pedagogical experiment was the need for a shift in the teacher–student model of interaction. To prove the benefits of semi-independent student's research, the article covers four research cases performed by 3rd year students at Moscow State Linguistic University under the guidance of their English grammar lecturers. While working at their respective tasks, the students tried their hand at a proper piece of scientific study and mastered their skills at a new research instrument. Corpus data obtained, they worked in cooperation with their lecturers to arrive at conclusions reviving textbook materials. Thus, the rules and tendencies verified were no longer offered to the class as ready-made answers. The results of the experiment foreground the positive effect of students' getting access to novel technologies transforming the traditional roles assigned in class.

Keywords: language corpus, linguistic education, grammar rule

Introduction

Recent technological developments have given a lease to the extensive use of massive data sets that are managed and processed only with the help of specially designed software, which has opened new horizons and has changed the surface of linguistic research and education introducing language corpora and corpus analysis as an indispensable tool for raising adequate linguistic knowledge.

Being an extensive collection of sampled texts, annotated with relevant linguistic information and suitable for computational (primarily quantitative) analysis, corpus offers previously unseen opportunities for a language researcher. The amounts of data that can be

---

[*] Corresponding author: smartletters@mail.ru
[†]Deceased.

computed at high speed convince that what was considered "impossible" and "lunatic" 40 years ago [1] has become not only possible and necessary, but imperative for conducting any quality language study. "The computer's ability to search, retrieve, sort and calculate the contents of vast corpora of text, and to do all these things at an immense speed, gives us the ability to *comprehend* and to *account for* the contents of such corpora in a way which was not dreamed of in the pre-computational era of corpus linguistics" [2, 3]. Numerous and diverse, the existing corpora of the English language can give an insight into many aspects of the language functioning and have already proved efficient for research into grammar structures [4, 5]. Thus, corpus awareness has become an important issue of language students' training.

Material and methods

The investigations described below address the grammar material studied by the students and aim at clarifying some debatable points. In their work, the students were instructed to use the most known and widely used Corpora found at [6]. The site presents a collection of sixteen corpora including samples of written and oral speech from texts of different genres and historical periods, balanced and annotated with relevant linguistic information. Apart from statistics on the frequency of occurrence of the requested symbols, be it words, word forms or word combinations of different lengths, and information about their discursive (genre), dialectal and historical features, the researcher can retrieve full contexts of their use (option Context) and compare their functioning presented in different corpora (option Compare). The Collocates option allows tracing the patterns in which a word occurs, by sorting the words to the left and/or right, while the contexts of its use in different collocations are found in the KWIC (Key Word in Context) section. The query syntax is universal for all the corpora on the site, specific information on query syntax can be found on the site in Help area.

Results

Herein we will present several corpus-based explorations carried out by the students of the English Language department at Moscow State Linguistic University (MSLU).

1.1   "You seem to have made a mistake…"

The research performed by Artem Kolygin (3rd year, MSLU) was devoted to the syntactical structures containing the verb *to seem*, specifically the complex sentence with the formal subject *it* and the subordinate clause introduced by the conjunctions (1) *that*, (2) *as if* and (3) *as though*:

E.g. (1) '*It **seems** <u>that</u> the village shop will have to close down.'* (Cambridge Dictionary Online [7])

*(2) 'It **seems** <u>as if</u> he wants everyone to feel sorry for him, but I don't.'* (ibid.)

*(3) 'It **seemed** <u>as though</u> time was standing still.'* (ibid.)

and a synonymous structure with a notional subject (Complex Subject):

E.g. (4) '*Everybody **seemed** to be ready, so we set off.'* (ibid.)

In all the four cases, the speaker expresses their uncertainty about the state of affairs, which is presented in the finite (1-3) or the non-finite (4) clause. However, the subject of sentences 1-3 is formal (introductory) and only functions to fill in the semantically empty

initial position, while in example 4 it is notional, referring to the agent of the action expressed by the infinitive.

The corpus-based comparison of the frequency of occurrence of these structures allowed the student to draw conclusions about the tendencies in their use. This knowledge is particularly noteworthy for Russian-speaking learners of English. Due to the absence of the syntactical structure resembling the English Complex Subject in Russian, the only option that they tend to use to express their doubt and guesswork is the complex sentence starting with *It seems (to me)…* Cf. Russian (*Mne*) *kazhetsya*.

The research that covered the total of 467,000 examples proved that the use of complex sentences with any conjunction is a lot less frequent in both regional variants of the English language:
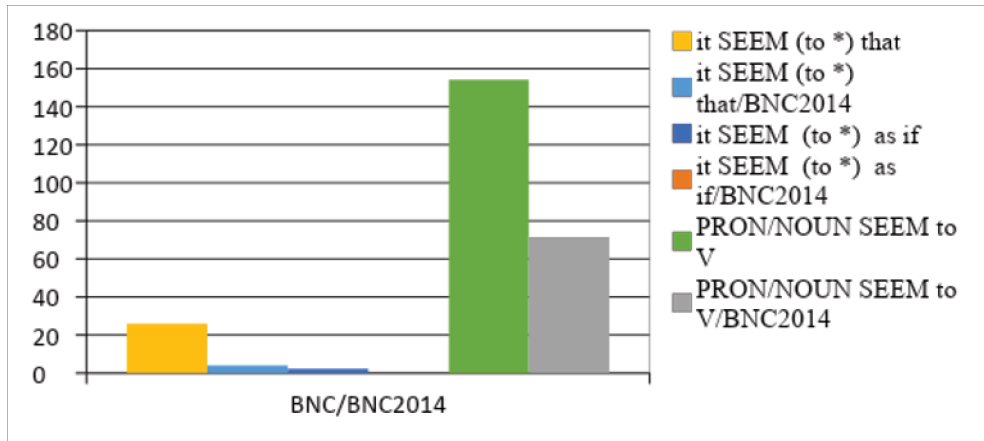
Fig. 1. Comparative frequency of occurrence of the structures with the verb *to seem* in BrE (British National Corpus (BNC) and BNC2014) [8].
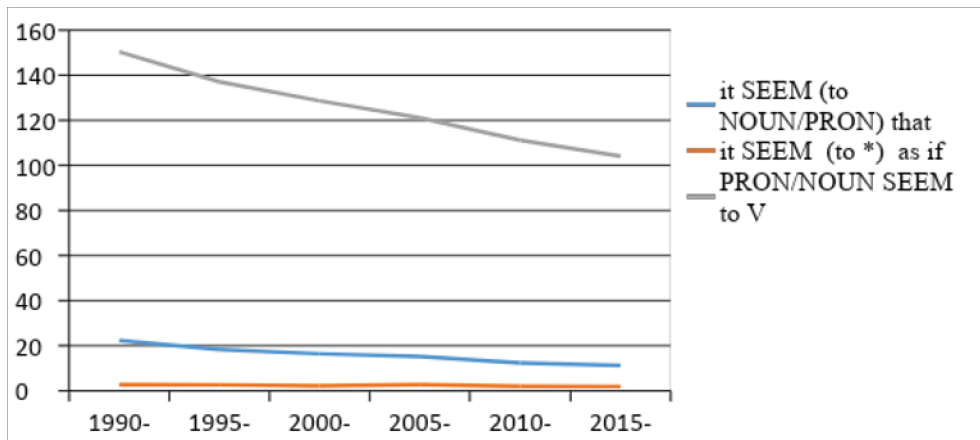
Fig. 2. Comparative frequency of occurrence of the structures with the verb *to seem* in AmE (Corpus of Contemporary American English (COCA)).

Judging by the numbers presented in the study, the structure with a notional subject is more acceptable in either variant. The reason for it may be the preference for semantic

completion of the utterance associated with the presence of the agent of the action represented by the grammatical subject [9].

1.2 **"Love and hate – what a beautiful combination…"**

The corpus-based study by Anara Botasheva (3rd year, MSLU) concerned the combinability of the verbs *to love* and *to hate*. Textbooks state that these verbs can be followed both by the infinitive and gerund [10], but there is no agreement between the authors either about the preference of this or that verbal or about the difference implied by the choice of either. Some grammars consider it the matter of regional preferences: while the gerund seems to be a likelier choice for the Brits, the infinitive is presumably preferred by the speakers of American English.

The query into BNC and COCA primarily showed that in modern use the gerund occurs slightly more often than the infinitive after the verb *to love* and almost equally often with it after the verb *to hate* both in BrE and AmE, so today the regional differences between these collocations are non-existent. It must be mentioned that the student had to compose the query so as to avoid the possible presence of the modal verb *would*, highly common in co-occurrence with these verbs and imposing the use of the infinitive, as in *I Would Love to Change the World* (Tom Jones) и *I would hate to see you go* (Billie Eilish, Copycat). As a way out, the query included not only the verbs *to love* и *to hate* as the predicate of a sentence, but the subject (noun or pronoun) as well; the resulting statistics were later summed up to show the total frequency of occurrence of the constructions
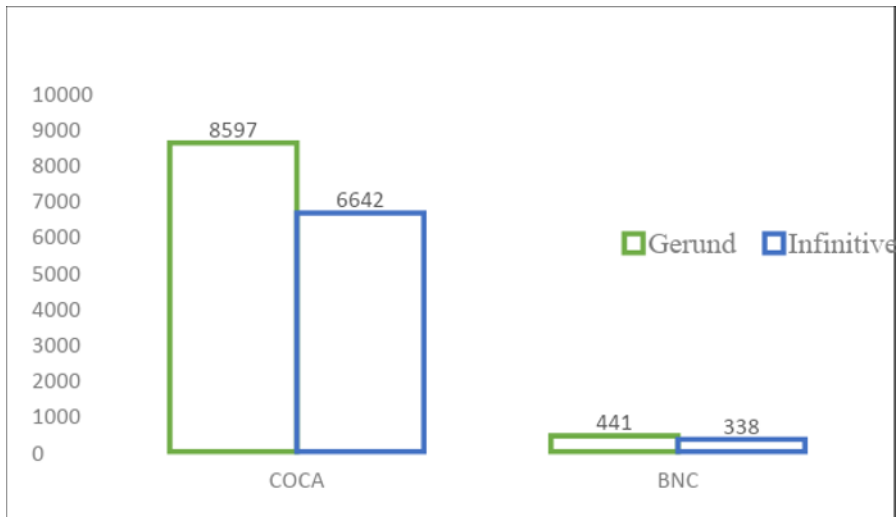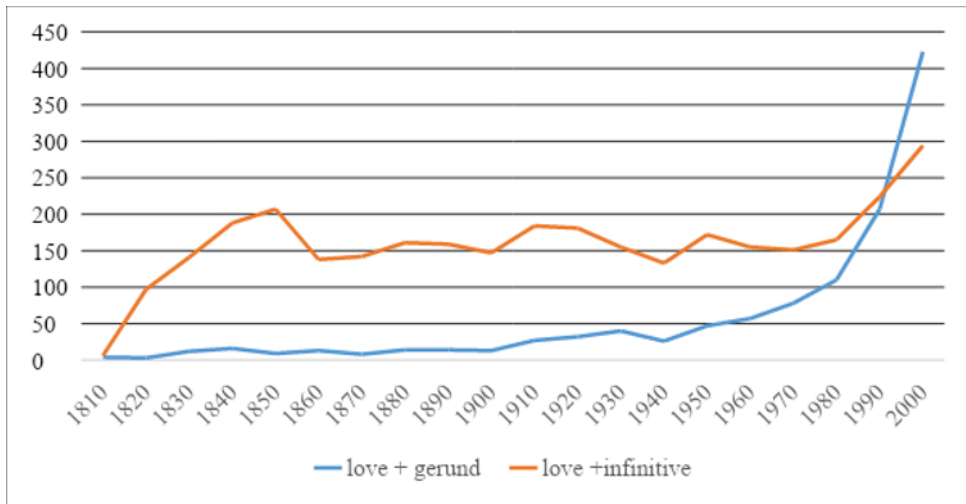
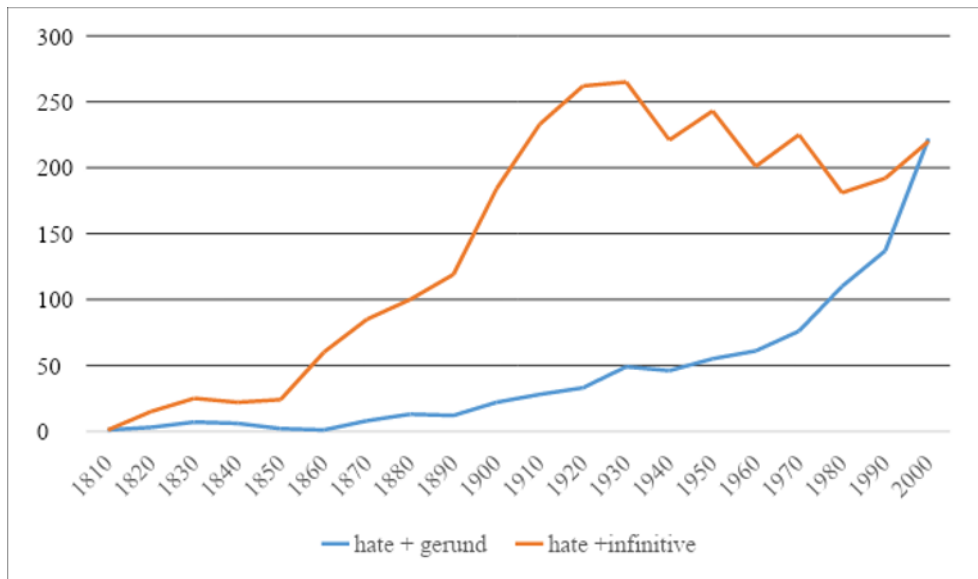**Fig. 3.** Comparative frequency of occurrence of the gerund and the infinitive after the verb *to love.*
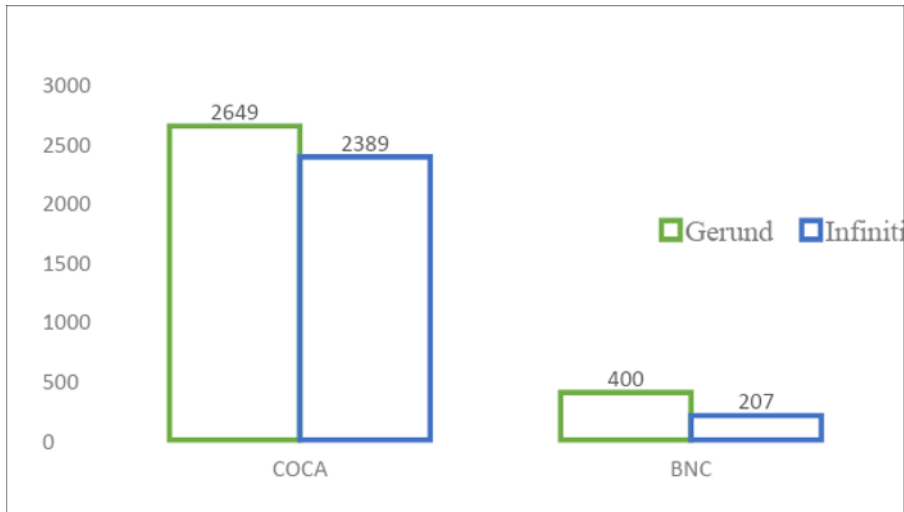
Fig. 4. Comparative frequency of occurrence of the gerund and the infinitive after the verb *to hate.*

However, a query into Corpus of Historical American English (COHA) showed that in AmE the preference of the verbal after the verb *to love* has changed only recently, the infinitive being the more expected choice until the 1980s.

Fig. 5. Changes in comparative frequency of occurrence of the gerund and the infinitive after the verb *to love* (COHA).

As for the verb *to hate*, the gerund slightly outnumbered the infinitive only in the 2000s.

SHS Web of Conferences **127**, 03006 (2021) https://doi.org/10.1051/shsconf/202112703006

*Methods of Teaching Foreign Languages 2.0: Real vs. Virtual 2021*

Fig. 6. Changes in comparative frequency of occurrence of the gerund and the infinitive after the verb *to hate* (COHA).

At the same time, a careful study of examples prompted that the choice of the non-finite form following these verbs is connected with semantic nuances: the use of the infinitive after the verb *to love* highlights habits and preferences while the gerund calls attention to the action that arouses the speaker's strong positive emotions. The verb *to hate* in combination with the infinitive acts as a synonym to the verb *to regret*, while the gerund implies feelings of antagonism or antipathy [11].

1.3   "This is a land of equal opportunities"

The research by Elizaveta Kopylova (3rd year, MSLU) similarly addressed an issue of combinability: it concerned the choice of the non-finite form after the noun *opportunity*. The practical relevance of the matter is conditioned by students often receiving controversial advice on the possible collocation. The query into COHA helped to resolve the problem: though the gerund and the infinitive were quite balanced choices up to the 1870s, the former has been rapidly falling out of use ever since, resulting in only sporadic appearance now.
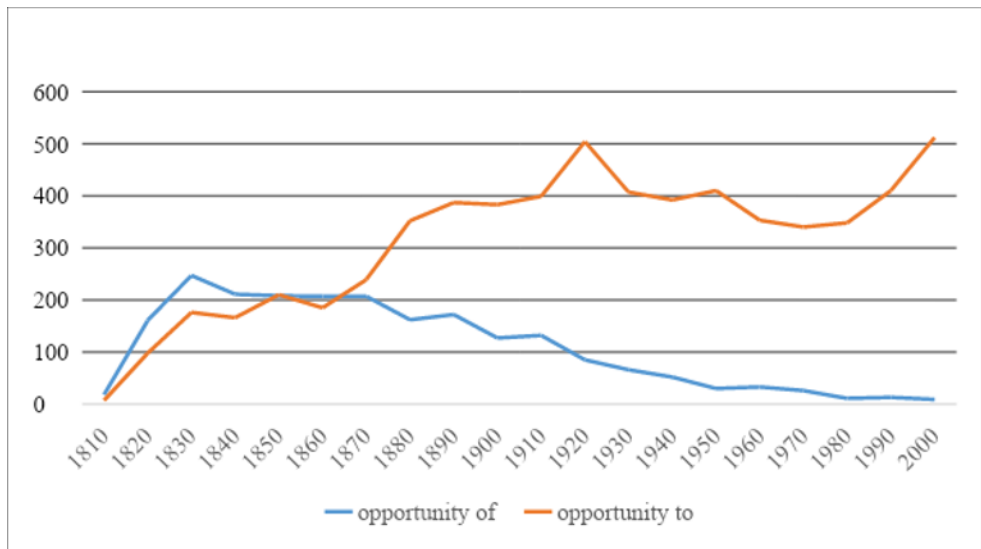
Fig. 7. Changes in comparative frequency of occurrence of the gerund and the infinitive after the noun *opportunity*.

So, according to the results of the study, the pattern *opportunity+gerund* is obsolete and in modern English only the infinitive is systematically found in this position [11].

1.4 Ask Google – Google it?

The investigation by Vladislav Orlov (3rd year, MSLU) dealt with conversion – a typically English word building means, which presupposes that a word (usually a noun) functions as another part of speech (usually a verb), acquiring its form-building features as well [12]. His task was to find out whether there are some historical and discursive preferences in the domain of conversion. To do it, he downloaded fragments of data from three corpora (COCA, COHA and NOW – News on the Web) – 2-3 mln of words–and created an algorithm that allowed to trace the intersections of the classes of nouns and verbs that were acknowledged to be cases of conversion. This was possible because all words in the corpus are marked as members of a certain word class. Naturally, the distinction between verb→noun and noun→verb conversion is impossible to make use of this method, but the manual analysis of examples proved that the latter pattern is much more common than the former. The corpora statistics unambiguously pointed at the existence of discursive preferences of this linguistic device: e.g. according to COCA, the rate of its use compared to "non-converted" words is 20.8% in media texts opposed to 16.3% in fictional and academic literature.

The statistics received from COHA was particularly curious: it showed that the frequency of occurrence of words that changed the part of speech has been constantly growing for the past 200 years mainly due to the increasing popularity of this device in mass media.
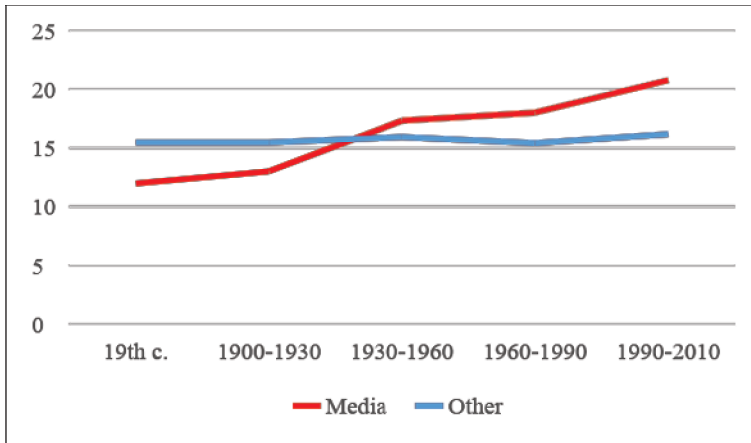
Fig. 8. Dynamics in the use of conversion (%) in texts of different genres in 19th-21st cc. (COHA).

Coupled with the results of the synchronous analysis this leads to the conclusion about the relevance of conversion to the needs of media style seeking brevity alongside profusion of expression. The use of "verbed" nouns serves this purpose to the fullest extent and contributes to the clear-cut and distinct media style.

Discussion

Obtaining statistics and visualizing them in tables, diagrams and charts take the linguistic exploration only half way. The other half means interpreting the data, and this requires the teacher-student collaboration of a new type. Usually, students receive grammar rules from the teacher or a textbook in a ready-made form, and their task is to simply follow the instructions. Meanwhile, a turn to corpus suggests a change of the classroom paradigm: it "entails a shift in the traditional division of roles between student and teacher, with the student now taking on more responsibility for his or her learning, and the teacher acting as a research director and research collaborator rather than transmitter of knowledge" [3, 13]. The problem-based approach to learning implying the search for information on real language use in corpus inspires students to draw their own conclusions with regard to the material studied and incites their intellectual activity [14], supported and guided by the teacher.

Another beneficial aspect of the use of corpus data in class is that students are exposed to authentic language: they can establish the patterns of naturally occurring speech, observe historical trends, regional and discursive variation. For example, students may discover that the self-reference *we*, a typical feature of business discourse, functions as a corporate *we* referring to an organization [15]. This makes a valuable contribution to their language awareness, still more precious for those who study a language out of the native language environment.

Conclusion

Introduction of hands-on corpus activities allows solving relevant educational tasks: verifying and intensifying grammar rules presented in textbooks as well as molding perfect linguistic and instrumental competences of a will-be linguist. Moreover, the students'

enthusiasm about using recent technologies and technology-proven validity of results prompts the conclusion that corpus analysis is a helpful flexible tool to promote more independent and conscious learning, and develop students' autonomous research skills, with the teacher ready to offer advice and expertise.

References

1. J.M. Sinclair, Corpus, Concordance, Collocation (Oxford University Press, Oxford, 1991)

2. G. Leech, *Corpora and Theories of Linguistic Performance*, in J. Svartvik (ed), Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, 105-122 (Mouton de Greyter, Berlin, New York, 1992)

3. Y.A. Breyer, Corpora in Language Teaching and Learning: Potential, Evaluation, Challenges. (Peter Lang, AG, 2011)

4. U. Römer, S. Skalicky, N. Ellis, Corpus Linguist Ling Theory, **16(2)**, 303-331 (2020). https://doi.org/10.1515/cllt-2016-0055

5. M. Sung, Corpus Linguist Ling Theory, **16(1)**, 189-214 (2020). https://doi.org/10.1515/cllt-2017-0002

6. English corpora. Accessed on: October 11, 2021. [Online]. Available: https://www.english-corpora.org

7. Cambridge Dictionary Online. Accessed on: October 11, 2021. [Online]. Available: https://dictionary.cambridge.org/grammar/british-grammar/seem?q=Seem

8. British National Corpus. Accessed on: October 11, 2021. [Online]. Available: https://www.english-corpora.org/bnc/

9. A.A. Kolygin, *Issledovanie sintaksicheskikh struktur s pomoshchyu korpusnogo analiza (na primere glagola to seem)* [Corpus-based analysis of syntactic structures (the case of the verb to seem)], in Collegium Linguisticum, Proceedings of MSLU Student Scientific Society, 25-26, Moscow, MSLU, 19–20 March (2020)

10. I.P. Krylova, E.M. Gordon, Grammatika sovremennogo angliiskogo yazyka. Prakticheskii Kurs [A Grammar of Present-Day English. A Practical Course] (University Book House, 2011)

11. E. Kopylova, A. Botasheva, *Diakhronicheskaya i zhanrovaya variativnost sochetaemosti glagolov "love" i "hate" i sushchestvitelnogo "opportunity"* [Diachronic and genre variation in the combinability of the verbs love and hate and the noun opportunity], in Collegium Linguisticum, Proceedings of MSLU Student Scientific Society, 26-27, Moscow, MSLU, 2020, 19–20 March (2020)

12. V. Orlov, *Polevoe issledovanie idioma sibirskikh seto: pervye itogi* [Field Study of Siberian Seto: First Results], in Collegium Linguisticum, Proceedings of MSLU Student Scientific Society, 37-38, Moscow, MSLU, 2020, 19–20 March (2020)

13. T. Johns, *Whence and whither classroom concordancing?* in T. Bongaerts, P. de Haan, S. Lobbe, H. Wekker (eds), Computer Application Language Learning, 9-27 (Foris, Dordrecht, 1988)

14. A. Mustajoki, *Rol korpusov v lingvisticheskikh issledovaniyakh i prepodavanii yazykov* [The Role of Corpora in Linguistic Research and Teaching], in The National Russian Corpus and Humanities Teaching: Proceedings of the International Scientific Conference, 58-60, Moscow, April 19-20 (2007)

SHS Web of Conferences **127**, 03006 (2021)

https://doi.org/10.1051/shsconf/202112703006

Methods of Teaching Foreign Languages 2.0: Real vs. Virtual 2021

15. N.E. Azizova, Sci Edu, **2(1)**, 266-271 (2021)