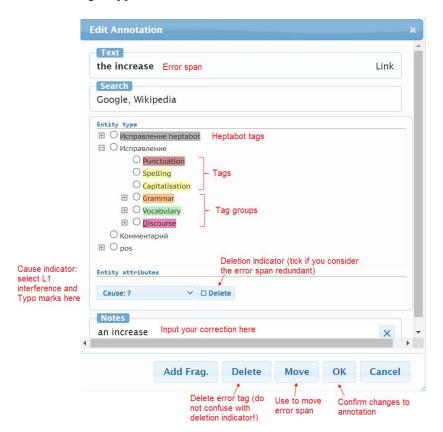
REALEC Tagging Manual

1. General principles

The best browser to access REALEC is Google Chrome. You can access the corpus at https://realec.org/index.xhtml#/exam/. To start annotating a text in the corpus, type in your login name and password in the upper right-hand pop-up window (write to REALEC administrator ektsann@gmail.com to get login and password). Click on the folder you need, and a collection of files opens. Essays in folders Exam_2014 – Exam_2019 were annotated manually. Essays in folders starting with Exam_2020 were initially annotated by the neural network, whose grey tags look different from the colour tags in REALEC error taxonomy. The work is different for manual annotation of essays from the former folders and for transforming grey tags into colour ones in the latter folders. Nevertheless, the principles of assigning tags will be the same, so the next section lists them before outlining the strategies for either type of annotation.

To select a REALEC error tag, open the annotation window by double-click on the error span and click on the plus next to **Исправление**. This will reveal the most general level of tags - **Punctuation, Spelling, Capitalisation, Grammar, Vocabulary,** and **Discourse**. You will see pluses next to the latter three tags, which means these tags have further subdivisions. The **Notes** section gives the corrected version for the error span. It should be as close as possible to the author's version, and the correction can contain only one suggestion. Separate words or the entire error span from **Text** can be copied to **Notes** if you select the necessary span and drag it down until a rectangle appears next to the cursor. See Annotation window in Figure 1.



2. Brief instructions for selecting REALEC tags

2.1. Annotation of errors with "Punctuation", "Spelling", and "Capitalisation" tags

These categories have no subdivision. "Punctuation" labels are dark red, while both "Spelling" and "Capitalisation" are yellow.

The error span for "Punctuation" tag will be:

- the introductory or parenthetical expression that has not been separated with a comma at the
 end or with two commas, at the beginning and at the end has to be copied in the Notes
 section with the comma after it or with commas before and after it;
- the construction (attribute or clause that precedes the main sentence) has to be copied in the
 Notes section with the commas before and after it, if they are present in the text;
 - (1) If people walk in the streets where all the houses are similar and dim their moods will also be ugly as well as the buildings > If people walk in the streets where all the houses are similar and dim, their moods will also be ugly as well as the buildings (if the clause follows the main sentence, no comma will be needed see next section);
- the conjunction or the conjunction with a comma, if a comma is missing before a conjunction or if there is a refundant comma before a conjunction, and the correction in **Notes** will have the same conjunction, correspondingly, with or without a comma.
 - (2) Overall, it can be clearly seen, that the percentage of fossil fuels energy in Sweden and France was declining throughout the period > Overall, it can be clearly seen that the percentage of fossil fuels energy in Sweden and France was declining throughout the period
- Dash with the word following it, if the dash needs to be deleted or corrected to a different mark.

See section 2.4. about NOT applying "Punctuation" tag for mistakes in Relative clauses.

Examples of mistakes to be tagged "Spelling":

(3) This type of help is widely-practiced > widely practiced and and supports the opinion of helping people globally; its > it's a waste of time; at the same time tha > that is why the cost does not improve the quality; meanwhile ordinary > ordinary houses remained safe (in the last two cases, the mistake is caused by a typo, so you will need to select Typo in the drop-down menu Cause).

A missing space between words will be tagged as "Spelling" with "Typo" selected as cause if it is a mechanical error, but "Typo" is not needed if the author confused two homophones (anyway - any way, all together - altogether).

If the wrong homophone has been chosen (*loose* instead of *lose*), it should be tagged with both "Spelling" and "Choice of lexical item".

"**Spelling**" should also be used if the author simply transliterated a Russian word (for example, *abonement*).

2.2. Annotation with tags in the group "Grammar"

In these group, there are two label types: orange, which correspond to the area of morphology, and blue for the errors in syntax.

Depending on the part of speech of the word with an error, an annotator has to look up tags in the following groups – "Determiners", "Verbs", "Nouns", "Prepositions", "Conjunctions", "Adjectives", "Adverbs", "Numerals", "Pronouns". All except "Prepositions" and "Conjunctions" have further subdivisions.

Mistakes involving words *another/the other/other(s)* are annotated with the general tag "**Determiners**":

The other > Another example, some hospitals in Moscow.

- "Articles" is the only subdivision of "Determiners". When using this tag, it is important to include the word the article is supposed to be related to:
 - (4) The given chart and table indicate the percentage of children of both sexes who are overweight and the percentage of children who do exercises on regular basis > a regular basis; with the exception of India, which has a lowest figure > the lowest figure of 35% this energy used.

However, the noun does not need to be included if it does not determine the choice of article: these cases include the usage of the superlative degree of adjectives and adverbs, as well as adjectives that require "the" when used before nouns (same, wrong, right, very, following, whole, main, principle, central, present, former, latter, necessary, opposite, previous, usual, so-called).

The following example should not be annotated with "Articles", but with "Redundant component in a clause or sentence", since the noun that the article would be related to is missing:

- (5) The analysis of figures show that slight grow was experienced by Sweden during the period between 1940 and 1990 from the > -- around 7,5% to approximately 15%.
- "Nouns" is a group that includes several tags. "Prepositional noun" is assigned if the choice of preposition is determined by the noun that precedes it:
 - (6) The graph depicts the **information of > information about** the energy from fossil fuels in five different counries from 1970 to 2015.
 - (7) What is more, in 1984 there was a sagnificant dicrease in the consumption of energy in Sweden but in that year in India there was a **rise of > rise in** it.

If a plural is formed incorrectly (think *childs* instead of *children*), you should choose the tag "*Noun number*", if a plural of an uncountable noun is formed (*advices*), choose "Countable/uncountable".

Example of usage of "Prepositions":

(8) On the one hand, some people think that the government must hepl those, who really need it **despite > regardless of** the motherland of the person.

Example of usage of numerals "Numerals":

(9) but over time the index raised to three third > three thirds of total percentage by 2015

The group with the largest amount of subdivision is "Verbs" – it includes "Tense", which is further divided into "Choice of tense" & "Tense form", "Voice," "Modals," "Verb patterns," "Infinitive constructions" и "Participle or gerund constructions".

Example of usage of "Choice of tense":

(10) The pick was in 1970 - four fifth (82%), while in 2015 it **fall > fell** to one fifth(20%).

"Voice" is used for mistakes in voice for verbal forms, as well as adjectives ending in -ed/-ing:

(11) Also, one particularly **interested > interesting** fact highlited by graph is that in 2015 the leader in fossil fuels energy was USA.

Example of usage for "Modals":

(12) In other countries such as United States, France and Sweden we could noticed > could notice a dipping.

To understand the usage of "Verb patterns", we need to consider several examples. The first two can be seen below:

- (13) But from my point of view, when we do this kind of collecting data, we just work on understanding of > understanding another person's opinion.
- (14) It is clear that approximately 70% of young people between 20 and 29 have ideal weight in both charts; however, situation for those who are in 60-69 age group is completely different, with more than a half in 2010 suffering with > suffering from obesity.
- (15) To conclude, I'd like to say that despite the fact the appearence of building is not the most important thing, we should **let architects to create > let architects create** something new and beautiful.

It is necessary to learn to distinguish the cases when the preposition is included in the verb's government model from cases when it forms a separate adverbial modifier, albeit syntactically connected to the verb. Errors concerning compatibility of auxiliary verbs and adjectives (*be*

proud of, be dependent on, etc.) are annotated with "Prepositional adjective", a subdivision of "Adjectives." Prepositions after verb forms can be annotated in several ways:

- (16) They shake with fear, their hands begin to sweat, and hearts try to jump **out from** > **out of** their chests. tagged "Prepositions" since the preposition out of (the correct variant) introduces an adverbial modifier which is not a necessary argument of to jump.
- (17) This leads to some difficulties in the distinction between the two concepts and this is where the main problem lies: how to distinguish among > distinguish between homonymy and polysemy? tagged "Verb patterns" since the verb distinguish has 3 government models:
- to distinguish smth/smb from smth/smb
- to distinguish between smth/smb and smth/smb
- to distinguish smth/smb among smth/smb

The first two models behave semantically similarly, both concerned with distinguishing between two objects, while the third one implies distinguishing a singular object among the whole group. When correcting the mistake in this sentence we can choose any of the first two models: if we choose the first one, we correct the error span to *«how to distinguish homonymy from polysemy»*; and if we choose the second, we correct it to *«how to distinguish between homonymy and polysemy.»* In this context the third model is inappropriate since the context implies only two classes.

Other examples of errors annotated with "Verb patterns":

- (18) How can I show it? I'm writing about > writing and cannot draw a picture in my essay.
- (19) Usually I write by > write poems. This mistake is more likely caused by the interference with the Russian expression «nuuy cmuxamu» it should be denoted by selecting "L1 Interference" in the Causes list.
- (20) Please introduce > introduce yourself.
- (21) If they want to wash > wash themselves in privacy, they have to wait for hours.

NB: It is important to remember that the English verbs *afford* and *feel* do not include a reflexive pronoun in their models, as opposed to their Russian equivalents *позволять себе* and *чувствовать себя*:

- (22) When I do not **feel myself > feel** well, I cannot travel across the city. This should also be denoted by selecting "L1 Interference".
- (23) I have always **regarded him to be> regarded him as** my friend.
- (24) The experts, however, consider it's not > do not consider it to be dangerous.
- (25) She was dropped > dropped out of university after the first set of exams.

- (26) Then she described me > described to me the days spent in this wonderful city.

 Mistakes in complex verbal predicates are also tagged with "Verb patterns":
- (27) If one day the majority of people **start uses >start using** sun energy for vehicles it will be great success.
- (28) Will Gardner, a founder of a law firm, informed his employee Alicia about this case and demanded to > demanded that she deal with it.

Example of usage of "Participle or gerund constructions":

What I mean here, is that **create > creating** modern and beautiful house is vitaly important for saving our culture.

2.3. Annotation of syntactic errors

Mistakes in syntax are usually classified by the type of construction where they are in. The tags are blue in our scheme, they also belong to "Grammar" group of tags, and they are the following: "Agreement" (for mistakes in number agreement between subject and predicate or noun and its demonstrative pronoun determiner), "Word order", "Relative clauses", "Parallel constructions" for mistakes involving using different forms in parallel constructions; "Negation" for inappropriate usage of double negative or absence of negative forms where needed; "Confusion of structures" for cases involving contamination between two different syntactical structure; and "Comparative construction" for mistakes such as these:

(29) Twice more > Twice as many people in this age group use LinkedIn. This mistake should also be marked with "L1 Interference".

Examples of mistakes relevant to the tag "Agreement":

- (30) These scyscrapes is > are an outstanding feature of Moscow, which looks as modern as United Arab Emirates. (The verb is selected as an error span, since agreement is disrupted with the choice of its form).
- (31) For instance, migrants from various countries might occupy the whole cities and state their own rules in **this > these** areas.

Example of mistake tagged with "Parallel constructions":

(32) Many people all over the world like to spend their free time walking around the city and enjoy > walking around the city and enjoying the beautiful landscapes of the buildings.

"Relative clauses" is used not only for punctuation mistakes in relative clauses (see above), but also for erroneous choice of conjunctions. If there is a redundant comma or the necessary comma is only absent **before** the relative clause, the conjunction is considered the error span, either with or without a comma, and if this happens at **the end** of the construction, the whole clause is considered the error span. It is to be pasted into **Notes**, and commas are to be added or removed as necessary. If the conjunction is erroneously chosen, it will be the only part of the error span.

- (33) For instance, there are many offices in China which > whose role is to be not as hazardous and useless for the nature as the other ones.
- (34) Moreover, with the help of various developments and various mindsets people combine the most useful characteristics and create products, which > which assist to the whole humanity.
- (35) Living in modern world, which is full of new technologies >, which is full of new technologies, it may become extremely complicated for parents to control their children.

Example of usage for "Confusion of structures":

(36) There is an opinion that **there is > it is** more important for building to be used for its purpose rather than look pretty and the architects don't need to make sense of it at all

2.4. Annotating lexical mistakes

The overwhelming majority of lexical mistakes made by the learners of English (green-colored tags in the group named "Vocabulary") can be divided into two types - mistakes in word choice and mistakes in word formation. Correspondingly, there are two categories of tags under "Vocabulary": "Word choice" and "Word formation". The former is further subdivided into "Choice of lexical item" and "Change, addition, deletion of a part of collocation". Here are some examples of mistakes annotated with the latter tag:

- (37) This solution is very common for both Europe and the USA, where the percentage of people having having own >their own cars is higher, so for Russia it's a bit less effective but nevertheless it may help a lot.
- (38) There is a widely common opinion that the way the building looks like does not matter and it just should **make its function > fulfil its function**.
- (39) In the modern world of today the jury is still on > the jury is still out on whether or not the government should aim to provide help for all the countries.

Example of usage for "Choice of lexical item":

Secondly, some buildings carry > carry out a function of historical monuments (phrasal verbs are not considered collocations, so the second tag is inappropriate).

Complex errors can be annotated with the general tag "Word choice":

(40) However, due to economic instability people lose their places in work industry > work places and they have to ask help from the government.

Word formation group of tags is used for mistakes, firstly, in choosing suffixes and prefixes or in the lack of them ("Derivational affix"), and secondly, for cases of choosing the wrong part of speech from a word family. This last tag — "Confusion of categories" — is the only tag in the group that describes the nature of the mistake and not that of the correction. In the following example there are two mistakes which were assigned "Confusion of categories" tag:

According to the graph, the vast majority of young **adulting > adults** tend to use Facebook or Instagram on a daily basis, with 87 and 53 percent respectively, **contrasting > in contrast** to the usage of LinkedIn at only about one fifth of this age group.

Other examples of "Confusion of categories":

- (41) For example, every school has to have comfortable fully equipped classrooms, a **sport > sports** gym and a cafeteria.
- (42) In the given graphs the information is presented on the topics of the **overweight** > **obesity** of girls and boys.
- I am agree > agree with the second statement (in this case, the author is mistaken the verb for an adjective under the influence of the Russian expression Я согласен, so it needs to be marked as L1 Interference).

If there is a mistake in one of the parts of a compound word containing a dash, the whole word needs to be selected as the error span. Nonexistent attempts at a compound word such as *alcodrinks* need to be tagged as "Word formation", not "Compound word", since the correct compound word does not exist.

Example of usage of "Derivational affixes":

From my perspective, it is quite **controversional** > **controversial** issue.

Cases when similarly sounding suffixes are confused (-tion/-sion, -ance/-ence) require a combination of "Derivational affixes" and "Spelling".

2.5. Annotating discourse mistakes

"Discourse" group of tags helps annotate mistakes affecting the logical structure of the text, mistakes in construction of its parts and connections between them. The principal subgroups of tags here are - "Coherence" with a subtag "Linking device", separate tags "Referential device" (used for mistakes in referential devices, their absence or superfluousness); "Inappropriate register" (for annotating expressions inadequate for academic writing, and sexism in language like using he, him, his when referring to people of both sexes); "Absence of component in a clause or sentence" (for constructions lacking a certain element and, thus, difficult or impossible to understand); "Redundant component in a clause or sentence" (for an element whose presence causes a syntactic defect or gets in the way of understanding the text); "Absence of the necessary explanation or detail" (used for syntactically correct and legible pieces of text that call for further explanation, for example, in the first sentence the author tells us that the graph reflects the results of the study conducted in three countries, yet in the second sentence he or she proceeds to analyse the data without specifying which country it applies to).

Examples of the discourse tag "Referential device":

(44) It is important to mention that the amount of female graduating students was always higher than of > than that of male graduates.

- (45) We have computers, smartphones, laptops, and **it >they** influence the way we live, study and work.
- (46) There is a statement that great Goethe, when he was doing different types of work, he>-- used spectacles with glass of different-colors.

In sentence (20) the subject is followed by a subordinate clause, which is followed by the predicate, so no reference is necessary. Erroneous usage of referential personal pronoun leads to the predicate having two subjects at once: 'Goethe' and 'he', so the latter was deleted. It is possible that the mistake was caused by the influence of the pronoun 'he' in clause 'when he was creating different types of work'.

Examples of using the "Inappropriate register" tag:

- Essays in REALEC corpora must be written with respect to the rules of academic writing. That is why words of the low register and slang are substituted with stylistically neutral equivalents, for example: babies > young children, very old people > the elderly, teens > teenagers (except for 'in their teens'), etc.
 - (47) It is like a basis for today's literature, which is characterized by simple vocabulary and **pretty** > **easily** understandable plot.
 - (48) Other groups show less distinct change: 70% among the **youngsters > young** and...
- One more example of assigning this tag is using the pronoun *he* as opposed to a gender-neutral pronoun:
 - (49) If a person is hungry and cold, he > he or she will not feel much admiration towards the beautiful streets of his > his or her city.
- That mistake was made because of interference with Russian, where the choice of gender for a referential pronoun masculine, feminine or neutral depends on grammatical gender of the noun or pronoun to which the reference is made. Modern written tradition of English requires to either assign both pronouns of feminine and masculine gender as a reference to words like 'a person' or use the gender-neutral singular *they*. These cases need to be marked as "L1 Interference".
- "Coherence" has "Linking device" as its subtag for cases when the linking device or parenthesis used in the text disrupts the logical structure, or when clauses or sentences lack a necessary link between them.

Examples of using "Linking device" tag:

(50) Working in the office obliges everyone to work at the same time and with the same rhythm, which can lead to the decrease in the efficiency of working. Secondly, not coming to the office saves a lot of time that people usually spend on the way to work.

Nevertheless > Besides, if more people started working from home, there would be fewer traffic jams, and this would be also better for ecology.

- (51) Nowadays, with constantly developing computer technologies, new opportunities appear and working conditions change, > and the whole world becomes a world of mobility and rapid adaptability.
- "Absence of a component in clause or sentence" tag is used when the element missing makes the syntactic construction incorrect.
 - (52) The only thing we have to notice is that the number of people who were overweight was a little smaller, while **appeared > there appeared** more people who were obese.
 - (53) The last group is people 65 years old and above **use** > **who use** social networks on a daily basis.
- "Absence of necessary explanation or detail" differs from the previous tag in that it is used for error spans that are syntactically and semantically correct but lack an element that was discussed in the previous sentences or is necessary for some other reasons.
 - (54) As we can see, there were no obese or even overweight respondents among young adults in 1950. The percentage of **people > people in this group** whose weight can be named ideal is about 70%.
- In example (51) it was necessary to clarify that the percentage of people with "ideal weight" is given among young men and women and not among all age groups.
- "Redundant component in clause or sentence" tag is more commonly used for the cases of unnecessary repetition or for redundant words that disrupt the syntactic or semantic structure of the clause or sentence without conveying any important information:
 - (55) To summarize all above, it can be said that in case of celebrities the media has to cover their private life mainly because there is a great demand for it from **both opposite** > **both** sides the stars and the public.
 - (56) Many children receive little attention from their parents **today** > -- because both parents work full time nowadays.
- In example (53) the annotator can choose which word is to be deleted *today* or *nowadays*, but both of them cannot remain in the sentence.

In the next example it is appropriate to use "Verb pattern", instead of "Redundant component in clause or sentence", since the redundancy of us is determined by the verb's valency:

- (57) Data **provides us > provides** information information that the lawest point of energy from fossil fuels in China was in 1970.
- (58) The cities will be losing their cultural identities and it will not be so exiting to visit new cities of countries > cities because they are all the same (the word countries needs to be deleted, as the sentence contains a pleonasm).

2.6. Special cases.

- 1. If you encounter a mistake in a construction, you need to deliberate whether the error span should include just the word or the whole construction. Thus, for example, a mistake in the choice of a preposition following a verb can turn out to be either a lexical mistake (put off with his egoism>put up with his egoism " Change, addition, deletion of a part of collocation"—put off is the error span), or a mistake in the verb's model (see on us > see us "Verb Pattern"—see on is the error span), or a prepositional mistake (seated on the armchair>seated in the armchair "Preposition"—on is the error span).
- 2. If a change in the error span necessitates a correction of the next word, this word needs to be included in the error span and its correct variant should be included in the correction in **Notes**. A separate error tag for this word is not needed.
- 3. One error span can include several annotations for example, "Spelling" and «Choice of lexical item" for a case where the author used *loose* instead of *lose*, or "Noun number" и "Articles" to correct the employee to employees. However, it is crucial that the correction in Notes for all annotations in the same error span is exactly the same.

2.7. Relations between tags

Sometimes after the changes have been made in one error span, the following element(s) may also need to be corrected. For this case, there is a mark denoting a relationship between two tags called **Dependent change**. Consider the example:

- (59) Everybody > They all ("Referential device") knows > know ("Agreement") it.

 To connect the two tags, click and hold on "Referential device" and drag an arrow to the second tag. A special window will pop up; select the option "Dependent change" and confirm by clicking "OK".
- 3. Strategy of work with errors identified and corrected by the neural network model. In analysing results produced by Heptabot, one has to scan every sentence for that have not been identified. When you have spotted such an error, select the error span and double click on it, and when the new annotation window opens, choose the correct tag from the list that opens if you press "+" before "Исправление" (not to be confused with "Исправление heptabot"!).

For every mistake spotted by the neural network, you first have to determine if the error span is correct. If it is, you will need to double click on the Heptabot tag and select the correct REALEC tag from the list that opens if you open "Исправление" (see above). This process will both apply the REALEC tag and delete the Heptabot tag. If the span is incorrect, you will need to double-click on the tag, press the "Move" button at the bottom of the annotation window, and

when the annotation window closes, select the correct error span, which will open the annotation window.

If you are making a new annotation, please do not forget to enter your suggested correction of the error span in the "Notes" section. If you wish to suggest a deletion, tick the "Delete" box in the Entity attributes section, which will add a red flag to the annotation. If you're selecting a tag for a mistake tagged by Heptabot, please check the suggested correction in the Note section and do not hesitate to correct it if you feel there is a better option.

When all the errors that you have been able to spot have been annotated, press the "**OK**" button to get back to the essay. You will see labels of different colours; hovering the mouse over the label allows you to see the correction suggested for it. A double click on the label will re-open the annotation window you have already been working with.

Ideally, at the end of your work, no Heptabot tags should be left in the text.

3.1. Difficult cases in analysing Heptabot suggestions

(1)



This case has both a temporary and permanent problem regarding its annotation. Firstly, the correction for this case is the, but it should follow the preposition for as opposed to substituting it, which is to be fixed in later updates. The permanent problem, however, is that the neural network inserts suggestions after the word as opposed to before, while the article always precedes the related word, which makes it necessary to change the error span from for to environment.

(2)

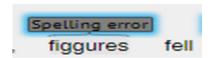


In addition to being misspelled, this word is used in the singular while the context suggests it needs to be plural. The neural network tag will still need to be substituted with the REALEC tag **Spelling**; however, the correction will need to be changed from business to businesses, and an additional **Noun number** tag with the same error span and correction will need to be added. To add an additional tag, press "OK" in the existing annotation window and open a new one by double clicking the word.

The bar chart illustrates the average time, which is spent

As explained in section 2.3., commas in relative clauses are tagged with **Relative clauses** instead of **Punctuation**, and the error span includes the comma (or the space where it should be) and the conjunction. In this case, you will need to first move the error span using the process described in section I, then select the Relative clauses tag in the annotation window for the new span, edit the **Notes** section to simply contain the conjunction which with no comma, and deselect the **Delete** box.

(4)



The tag itself is correct, however, the neural network suggests an irrelevant correction: fertility rates instead of figures. Do not forget to check the suggestion and edit the Notes section if necessary.

(5)

Sometimes you will encounter false positives and see the network flag correct words and suggest unnecessary corrections (for instance, replacing spent with spend when the text is describing past events). These tags should simply be deleted.

3.2. Automated transfer of Heptabot error tags to REALEC error tags