ERROR ANNOTATION IN LEARNER CORPORA: TOOLS AND APPLICATIONS IN ENGLISH AND ITALIAN

OLGA VINOGRADOVA, NIKITA LOGIN, IVAN TORUBAROV
RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS, MOSCOW

TALC 2018
13th Teaching and Language Corpora Conference
Wed 18 - Sat 21 July, 2018 - Faculty of Education, University of Cambridge
REALEC AND ITS ANNOTATION PRACTICES

Olga Vinogradova,
olgavinogor@gmail.com

TALC 2018
13th Teaching and Language Corpora Conference
Wed 18 - Sat 21 July, 2018 - Faculty of Education, University of Cambridge
VIDEO TUTORIALS

https://drive.google.com/open?id=1Q84ejBeZP5Hqs0t9pxTECWX3dM7Wv7h7

TALC 2018
13th Teaching and Language Corpora Conference
Wed 18 - Sat 21 July, 2018 - Faculty of Education, University of Cambridge
IMPORTANT LINKS

The main corpus: [http://realec.org](http://realec.org)

The section for training annotators: [http://realec.org/hse/index.xhtml#/Annotation_training/test_2/](http://realec.org/hse/index.xhtml#/Annotation_training/test_2/)

The annotation manual in English: [http://realec.org/tagging_instructions_eng.pdf](http://realec.org/tagging_instructions_eng.pdf)
GENERAL PRINCIPLES

- The best browser to access REALEC is Google Chrome.

- To start annotating a text in the corpus, type in your login name and password in the upper right-hand pop-up window.

- When you have spotted an error, double click on the error span. A new annotation window will open on your screen with the error span in the section above, the classification scheme with all error tags in the central

- To choose an appropriate tag for the error you are annotating, click on a plus in front of the first tag, and you will see the following tags of the highest level: “Punctuation”, “Spelling”, "Capitalisation", “Grammar”, “Vocabulary” and “Discourse”. A plus in front of any tag means there is a further division of this tag, and. You will see a plus in front of the last three tags on the list, and a minus for other tags - “Punctuation” or "Capitalisation" or “Spelling”.

TALC 2018
13th Teaching and Language Corpora Conference
Wed 18 - Sat 21 July, 2018 - Faculty of Education, University of Cambridge
Nowadays people can face the unprecedented problem of the high level of crimes that are committed by the young people.

To my mind, the main reason for such tendency is a freedom in Internet, where it is not very likely to find a censorship on most of the websites and also computer games.

To overcome this parents can restrict the access for children to the Internet.

First of all, parents can pay money for the specialist, who will install parents control on the computer, so children would not have any opportunity to visit webpages with specific content.

The positive effect from this is that the level of children aggression would really decline, since they would not be able to see violence and crimes.

However, there is no guarantee that they would not be able to watch it in their friend’s house.

Moreover, schools can try to make children to understand the consequences of bad behavior.

The idea is that schools should organize special lessons, where teachers will tell children about the rules and morality of our society.

Children are the ones, who perceive information very well, so they would grow up with the knowledge of etalon behavior.

On the other hand, this method can be not successful, since there is possibility of making the situation worse by constant reminding how to live.

All in all, I suppose, that the problem of youths crimes caused by violence is not going to be solved in nearly future because of the many aspects, though I think that the parents control is the most effective method today.
The article for young people should be indefinite, as it is about young people in general, and for a count plural noun the indefinite article is zero, so we highlight the entire noun group as the error span, choose Articles tag, and copy the noun and the adjective without the article as the correction in the Notes section below.
In sentence 2, we can see that the author of the essay does have serious problems with articles:

2 - with *such* the article has to be indefinite, and for a count noun it should be *a*: *such a tendency*;

3 - an abstract noun such as *freedom* has to have a zero article for Indefinite;

4 - in the majority of its uses, *internet* is used with the definite article *the*;

5 - one more abstract noun - *censorship* - has to be used with a zero article.
6 - the preposition for *internet* is wrong as well; it has to be *on*;

7 - , *where it is not very likely to find a censorship on most of the websites* is a non-defining relative clause that has a comma at the beginning, but no comma at the end of it, which confuses the reader into thinking that it is “censorship and computer games” when it actually is “freedom on the Internet and computer games.” The corresponding tag - “Non-defining relative clause” - is in Grammar, in its blue section (which is Syntax), under Attributes, further under Relative clauses.
In sentence 3,

8 - there is a punctuation mark - comma - missing for the infinitive of purpose at the beginning of the sentence;

9, 10 - two words are misspelled: parents > parents; accept > access; tag Spelling has to be selected

It may be discussed whether the latter is just a spelling mistake or a vocabulary mistake which has to be tagged with Choice of lexical item.

11 - the annotator may be concerned whether the access for the children should be changed for a much easier possessive form: children’s access. If so, it is to be tagged by Confusion of structures, another blue tag under Grammar.
In sentence 4,

12 - there is a mistake in the preposition to introduce an addressee - *for > to*, the tag we need is **Prepositions**;

13 - There is one more mistake in the punctuation of a relative clause, this time - redundant comma introducing a defining relative clause, so again, the corresponding tag is **Defining relative clause**, but what do we have to take as an error span?

14 - there is a subtle imperfection in the linking device between the two clauses: *so* has to be changed for *so that*.

TALC 2018
13th Teaching and Language Corpora Conference
Wed 18 - Sat 21 July, 2018 - Faculty of Education, University of Cambridge
In sentence 5,

15 - 17 - there are two forms would that do not make much sense either as a future in the past form, or as a modal, or as a subjunctive; but will instead of would makes sense, so we suggest **Choice of tense** tag on the following error spans: *would really decline > will really decline, and would not be > will not be.*

In REALEC one can tag not only the mistake itself, but also its relation to some other tag(s). The last two tags discussed are in exactly the same syntactic constructions in this sentence, so we introduce a relation between them - **Parallel construction**.

18 - plural noun *children* was used instead of its singular form, and there is a corresponding tag for it - **Noun as an attribute**.
19 - Sentence 6 has the same use of *would* as in the previous sentence, so again, we apply **Choice of tense** tag.

20 - **Sentence 7** covers the wrong pattern of the causative verb MAKE - *make children to understand* - it has to be annotated with **Verb + Object/Addressee + Bare Infinitive** tag, one of the 20 entries in **Verb pattern** group of tags:, and the correction has to be for *make children understand*.

21 - **Sentence 8** has the same mistake in relative clauses as in sentence 4 - redundant comma introducing a defining relative clause, and we choose the same tag - **Defining relative clause** - for the same kind of an error span - a comma and *where*.

TALC 2018
13th Teaching and Language Corpora Conference
Wed 18 - Sat 21 July, 2018 - Faculty of Education, University of Cambridge
In sentence 9,

22 - construction *are the ones, who* is clearly redundant, makes the sentence clumsy and should be tagged as **Redundant component in a clause or sentence**. In this case no correction is needed in **Notes**, but we tick **Delete** window, and the little red flag in the label appears;

23 - another tag **Choice of tense** similar to the one in **sentences 5 and 6**: *would > will*;

24 - *etalon* - the word exists in English, but in some special uses, and it is a mistake caused by Russian intuition, as in Russian exactly the same calque can be used in this context. We suggest changing it for *appropriate*, using the vocabulary tag **Choice of lexical item**. **REALEC** tagging system allows to reflect the cause of an error.
In sentence 10

25 - _be not_ after the modal is awkward, when the use of such lexeme as _fail_ is a solution to the problem. This has to be tagged with **Word choice** tag;

26 - a vocabulary mistake of choosing a wrong suffix in the word word _reminding_, which has to be corrected for _reminder_ with **Formational suffix** tag;

27 - two more uses of articles have to be corrected: _possibility - a possibility, constant reminding - a constant reminder_. **Note** that when error spans overlap - we have to ‘synchronize’ the corrections

28, 29 - two vocabulary errors in sentence 11: the first tag marks the change from _youths crimes_ to _juvenile crime_, and then the author confused adverb _nearly_ and adjective _near_. It has to be tagged with **Confusion of Categories** tag in **Word Formation** group of tags;

30 - one more incorrect use of Determiners: _the many aspects - many aspects_. Correspondingly, we need **Determiners** tag.

TALC 2018
13th Teaching and Language Corpora Conference
Wed 18 - Sat 21 July, 2018 - Faculty of Education, University of Cambridge
WELL DONE! GOOD WORK!
THANK You
IN APPRECIATION OF YOU, AND ALL THAT YOU DO.