Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2016"

Moscow, June 1–4, 2016

ЗНАЧЕНИЕ И ПРИМЕНЕНИЕ ЭКСПЕРТНОЙ АННОТАЦИИ ОШИБОК В КОРПУСЕ АНГЛОЯЗЫЧНЫХ УЧЕБНЫХ ТЕКСТОВ

Виноградова О. И. (olgavinogr@gmail.com)

НИУ ВШЭ, Москва, Россия

Ключевые слова: учебные корпуса; аннотирование; корпусные исследования; компьютерные инструменты

THE ROLE AND APPLICATIONS OF EXPERT ERROR ANNOTATION IN A CORPUS OF ENGLISH LEARNER TEXTS

Vinogradova O. I. (olgavinogr@gmail.com)

Research University Higher School of Economics, Moscow, Russia[1]

The paper presents the rationale for the decisions that were taken in the set-up and further development of a learner corpus of student texts written in English by Russian learners of English, the only Russian learner corpus in the open access. The tool of manual expert annotation is in the focus of the present observations, and after introducing categorization of errors applied in annotation the complicated cases that arose in annotation practices have been looked into followed by comparison of the annotation statistics over the three stages in the corpus development. For that purpose, texts annotated by different groups of participants in the process of two experiments were used to spot the problematic areas in annotation. The main pedagogical applications of the learner corpus in teaching EFL—the

[1] The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2016.

opportunities to create automated training exercises and placement and progress tests custom-made for specific groups of students—are outlined in the concluding part of the paper.

Key words: learner corpora; annotation; corpus research; computational tools

A learner corpus is a systematic computerized collection of texts that are written and/or oral productions of language learners. As all other corpora, a learner corpus is usually provided with convenient means of browsing and search options, with a system for marking the texts for pedagogical and/or research purposes, and ideally with additional visualisation of statistical processing of the search results. The first researches in this area of computational linguistics date back to late 80s—early 90s, and the main achievements in the area have been well reviewed in the collection edited by Granger, Gilquin and Meunier 2013. The main point of interest for linguists working on the development of a learner corpus is the choice of annotation. Learner corpora are usually *error-tagged*, which means that spelling, lexical, and grammatical errors in the texts have been outlined with the help of a standardised system of error tags. The exhaustive list of important references for the discussion of the use of annotation in learner corpora can be seen in Pustejovsky and Stubbs (2013) and Wilcock (2009). The following researchers wrote on approaches to annotation in different learner corpora: Granger S. (2003), Hovy and Lavid (2010), and on the decisions concerning the choice of annotation systems, see Shtindlova et al. (2014), Lee, Yan Yeung, Zeldes, Reznicek, Lüdeling, and Webster 2014, Glaznieks et al. (2014), and many others.

This article provides rationale for the decisions on corpus annotation taken in setting up one of the first Russian learner corpora and to our knowledge the only learner corpus of English student texts in the open access: it is free to search in and freely downloadable. The name of the corpus, REALEC, stands for Russian Error-Annotated English Learner Corpus and its texts are now available at http://realec.org and at http://realec.org/hse/#/data_4_staff. The focus will be placed on evaluating how the chosen tools and the annotation workflow affect the results of annotation. It concludes by discussing the prospects of how manual expert tagging in this particular corpus can be used in creating a few pedagogical and research applications.

The corpus now comprises almost 3,400 pieces of students' writing (with about 838,000 word tokens), of which essays written in preparation for IELTS and during the exam of the type make up the main part. It was initially set up as a pedagogical tool for EFL instructors who teach a course of general English, which includes preparation for IELTS, and for professors teaching Academic Writing in English. The initial goals were to provide those instructors with the tool for marking written works submitted by their students, as well as to give instructors the opportunities to carry out their independent research, and at the same time to provide students with the easy means to see which errors prevail in their writing. To satisfy these three areas of need, expert error annotation was designed the BRAT platform[2] (see Hovy (2015), p. 5 on growing popularity of BRAT).

2 Stenetorp et al. (2012)

REALEC at the present time has a well developed system of hierarchical tags to mark the errors, and these tags are shown above the text as labels in different colours along with suggestions on how to correct them. REALEC error annotation scheme consists of four layers: error type, error cause, linguistic 'damage' caused by the error and the impact of the error on general understanding of the text. The first of the annotation layers is the main source of knowledge about the mistake a particular student has made, so the paper only deals with this layer of annotation process, and the term *'annotation'* will be reserved for assigning tags that specify error type. The scheme includes 151 categories organized into a tree-like structure presented in Figure 1.

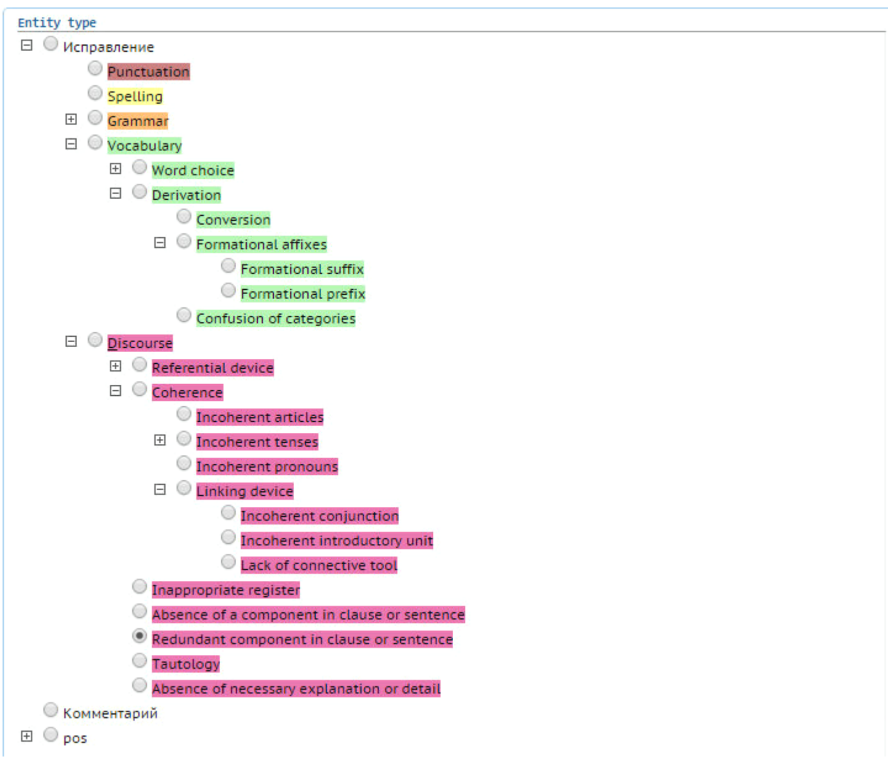

Fig. 1. Outline of the error-tagging scheme in REALEC[3]

In REALEC annotation, the following two important principles are observed: first, annotators mark as error spans only the areas of the text with clearly identified mistakes, and, second, they choose the most specific tag available in the scheme for

[3] As can be seen from a "+" in the little window, some error tags—namely, all of the *Grammar* area, *Word Choice* in *Vocabulary* area, and two tags in the *Discourse* area—*Referential device* and *Incoherent tenses*—are further subdivided into classes and subclasses of specific tags not demonstrated in Figure 1.

the error they have spotted, with the exception of some special cases, when they can assign one of the more general tags.

Expert annotation in a learner corpus has to be continuously evaluated. On the question of the comparative evaluation of expert manual annotation and automated error annotation, there are three important points:

- Learner corpora are, they are usually proprietary and often cannot be shared (Chodorow, et al. (2010) and Chodorow et al. (2014)). On the contrary, REALEC, as was mentioned above, is open for research.
- Learner corpora are as a rule expensive to annotate manually, and any alternative to time-consuming expert annotation has to be applied and tested. Some industrial applications have been reviewed by Chodorow et al. (2010) and Sorokin & Forsyth (2008), but at present they do not seem to give a valid alternative to manual pedagogical endeavours.
- The last two decades have seen an explosion in the development of NLP tools that aim to detect and correct errors made by learners of English as a Second Language (ESL) or English as a Foreign Language (EFL), so to meet this growing need, annotation schemes have to be built into the approach that combines automated detection of simpler errors with expert annotation of more sophisticated ones. An approximation to such a system can, for instance, be seen in the CEA = computer-aided error analysis (Diez-Negrillo & Fernandez-Dominguez (2006)); it is also presented in Yannakoudakis (2013). In the long run, there will arise the possibility of building a system that models human behaviour in the process of reading and making judgements about the value of someone's writing.
- When learner corpora are to be used to investigate learning process, high-quality corpus annotations as a basis for precise analyses are of great importance, and primarily it implies minimising the number of annotation errors on each individual level (Glaznieks et al. (2014)). That is why annotation schemes are always subject to scrutiny in the process of using a learner corpus, and as an example of this, Bayerl (2008) illustrates, on the one hand, various forms of 'annotator drift' as annotators get tired over time, and on the other, how their mutual agreement levels change over time during the exploitation of the corpus. This is precisely the area of the current research interests.

To check how the level of precision of manual annotation affects annotator agreement, we looked into the results of an annotator agreement experiment carried out in REALEC in 2015. There were 10 annotators involved—one leading the experiment, three English instructors familiar with the annotation practices, another three without any exposure to the annotation process in REALEC, and three more—students in computer linguistics proficient in English. All the participants were instructed on tagging practices at the beginning of the experiment and were given 30 student essays 150–300 words each with error spans outlined by the leader of the experiment, so that they had spot the error in the outlined areas and to look for the appropriate tag, or take off the mark if they did not see any mistake. The results—thirty texts tagged by ten annotated—were then subjected to two stages of research.

The first stage dealt with the procedure to calculate inter-rater agreement. The standard procedure is to use Krippendorff's alpha (further KA) (Krippendorff (2007);

Hayes and Krippendorff (2007), Krippendorff (2012)), Cohen's kappa (Cohen 1960), which corrects for chance agreement between two people, or Fleiss's kappa (see Passonneau (1997) for explanation on why precision and recall metrics are not feasible for our task). The goal of achieving a decent agreement among human annotators is difficult even for such an algorithm-prone system as grammar errors in a narrow subarea (see, for example, Bryant & Hwee Tu Ng 2015). Full agreement is almost never possible with any non-trivial annotation task, but the extent of agreement is still an important index of how reliable the adopted annotation method is.

Our 2015 experiment to check the rate of agreement among annotators was reported at the 8th International Corpus Linguistics Conference in Lancaster (Kutuzov, Kuznenko, Vinogradova (2015)), so I will only briefly state the results here. There were 2128 error category assignments in total involved. A topical question was how to apply KA in view of the hierarchical nature of our annotation scheme, and we did it by transforming our nominal scale of tags into an interval one. To explain, grammar errors differ one from another, but they are even more different from discourse errors. We assigned digital representations, or 'coefficients', to our error categories according to our intuitive knowledge of which categories are closer, so that tags belonging to closely related categories were assigned closer values. For the five macro-categories in REALEC, we assigned specific digital representations to subcategories. For example, the morphological part of macro-category Grammar is further divided into POS subcategories of Verb, Noun, etc. These tags are assigned different digital representations ("1", "4", "7", etc), whereas tags deeper down the hierarchy are assigned the same values as the upper ones. Between macro-categories we made 'gaps' 50 points wide. In the next level of the annotation scheme, we went down to the third-level subcategories (for example, Tense, Voice, Modals, etc). The same principle gave us the way to compute Krippendorff's alpha as if annotators had assigned interval digital values, and not nominal tags. As a result, we got Krippendorff's alpha = 0.57 for the second level annotation (tags like Noun, Verb, Word choice, Tautology, etc), even higher than at the upper level. The third level annotation had agreement rate equal to 0.55. Computing KA for the second and the third annotation levels as nominal categories (binary distance) gave only 0.5 and 0.4 correspondingly. The resulting index was satisfactory (KA = 0.57).

At the second stage, which has not been presented in a report or paper yet, the texts annotated in the experiment were used to research the cases of and spot the reasons for lack of annotators' agreement. I compared the results of each participant in each of the three groups of annotators with the results of all participants from two other groups, and then calculated the average values for each three participants of "EFL instructor familiar with REALEC/English student or instructor unfamiliar with annotation/computer linguist" type. Fig. 2 shows the statistics for the average "threesome".

The source data for this graph represents the average figures shown in this experiment, namely, 33 error spans per text on average initially outlined, of which in 6 on average annotators did not find a mistake and thus did not assign any tags, and among the 27 tags where some tags were assigned annotators agree on average in 17 errors and disagree on 10. The extent of agreement in this case can be different—annotators can agree in both the tag and the correction, or in one of them only. The following three examples illustrate it.

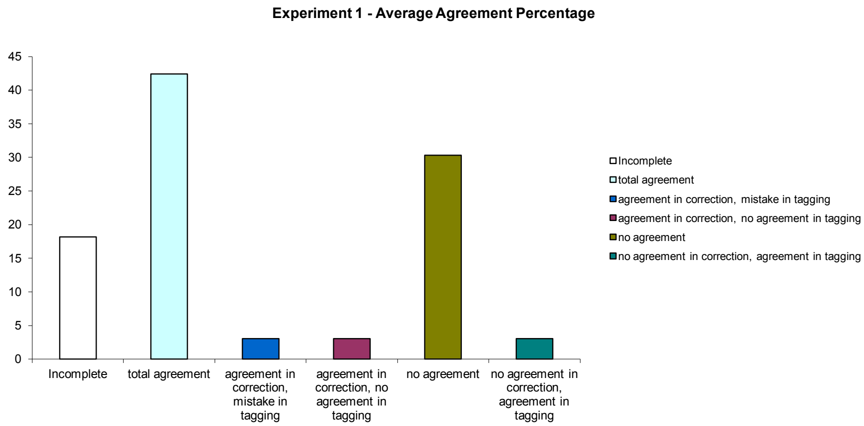

Fig. 2. Variation of agreement figures (average percentage) in agreement experiment

(1) *twice lucky> twice as lucky* (text 11) the same correction, different tags:
"Absence of certain component" (a vocabulary tag) 1 annotator
"Numerical comparison"—2 annotators
"Comparative degree of adverbs"—2 annotators—wrong tag!
"Prepositions"—1 annotator—wrong tag!
"Absense of a component in clause or sentence"—(a discourse tag) 1 annotator

(2) —*twice lucky> double lucky* (text 11) different corrections, different tags ("Vocabulary"—1 annotator)

(3) *And there was the same situation in 2001 with only a few variations in five cities* (text 3) the same tags, different corrections:
all annotators used a tag "Standard word order" and some discourse tag to change *cities* for *provinces*, as well as the tag "Preposition" to change *in* for *for* or *among,* and besides one more discourse tag—"Coherence"—to show the need for a change in the construction. Nevertheless, the resulting corrections were different:

>The same situation was in 2001, only with a few variations in five provinces
>And the situation was the same in 2001 with only a few differences for some provinces
>The same situation was in 2001 with only a few variations among five provinces
>The same situation was in 2001, only there were a few variations for the five provinces

In 2016, our goal was to trace the effect of changes that have taken place in our work over three years of active annotation practices. For this purpose, we collected data on the use of annotation tags in the following three areas of REALEC:

1. the initial student texts (essays, paragraphs and texts written in Academic Writing course, theses) collected over the first year of the corpus and tagged

by a group of students—participants of the research seminar (below referred to as ESL); the total of 1239 texts with 361240 tokens of error annotation;

2. IELTS-type essays from different departments of the Higher School of Economics dating back to 2014–2015 academic year and annotated by students in the Bachelor's course in linguistics at the HSE as their summer practical work (below referred to as IELTS); the total of 1941 texts with 433523 tokens of error annotation;
3. essays written in preparation for IELTS-type examination by students of one EFL instructor and annotated by students themselves or in peer tagging under the supervision of their instructor (below referred to as *current subcorpus* and labeled as 2ndYear 2015–2016); the total of 218 texts with 43181 tokens of error annotation.

In each part of the corpus, we collected data on the use of specific tags labeling student errors, and separately—on the use of highest-level general tags used by annotators. As stated above, the tag to be assigned has to be as specific as possible, and a higher-level (more general) tag can be used in one of the two cases—when there is no further division (for example, there no "Singular" or "Plural" tags for nouns—we only have a more general "Noun number" tag), or when the use of one more general tag simplifies the use of three or more specific tags of the same level. The example of the latter case is the following:

(4) *The almost equal number of increasing international graduates was observed…* *>The almost equal increase in the percentage of international graduates was observed…* (text 6)

An annotator can either use three specific discourse tags to show the errors made—"Coherence", for the change of *number* for the word *percentage*, the same tag for the change from *increasing*, and "Absence" of a component in a clause or sentence to add preposition of to the combination *increase in the percentage*, or choose to use one general tag—"Discourse" to signify the overall change.

Figs. 3–5 below demonstrate the variation in annotation statistics in three areas of REALEC:

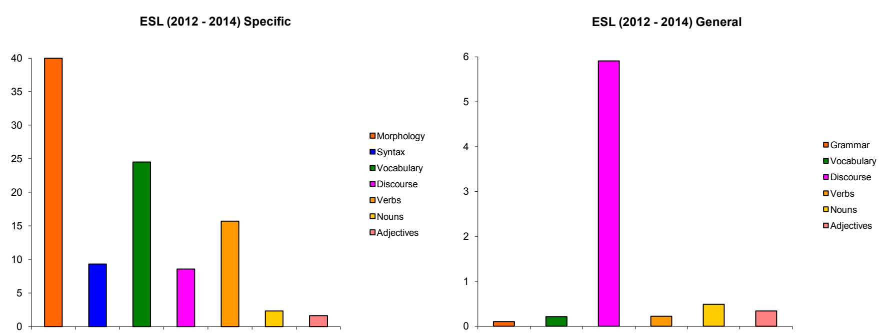

Fig. 3. Variation in the use of error tags in ESL, the initial learner corpus

It can be concluded from the graph on the right that general tag DISCOURSE was applied to the overwhelming majority of cases when annotators could not classify errors as grammar or vocabulary, and also that there was insufficient subdivision of discourse errors. Correspondingly, we worked towards eliminating these deficiencies by adding more discourse tags and working out specific approaches to annotating discourse errors. As a result, in the more recent addition to the corpus the distribution of tags assigned by annotators is more even:

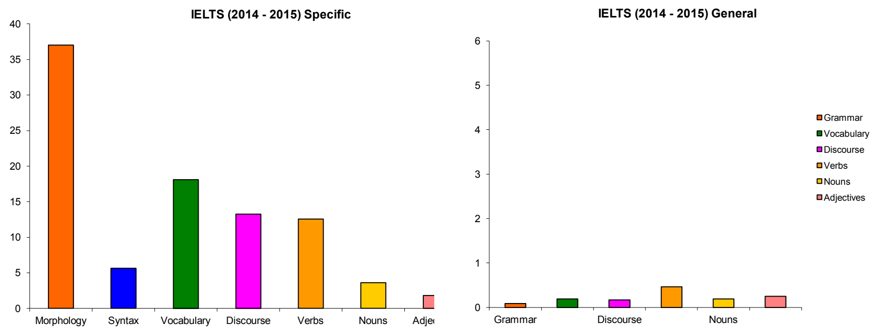

Fig. 4. Variation in the use of error tags in IELTS (the collection of examination essays in REALEC)

And finally, in the most recent texts added to REALEC—the essays written in 2015–2016 in preparation for their IELTS examination by the current students, who annotate themselves the errors that their instructors outline for them—there is only one case when a general tag was applied—it is the example very similar to the one discussed in (4) above:

(5) *It should be noted that the poorest group of poor people spends less on petrol—nearly 4 percent>It should be noted that the percentage of money spent on petrol by the poorest group of poor people in the two countries is very different.*

Instead of assigning three discourse tags—"Tautology" (because the same figure for the same group was given in the previous sentence), "Absence of the necessary information or detail" for the need to add in which country/countries, and "Coherence" for the need to talk about the difference for the two countries—the annotator decided to assign just one general tag—"Discourse," and for this single example of the use of high-level tag no graph on the right is presented in Fig. 5.

To observe more inter-rater differences in REALEC annotation practices, we carried out the experiment recently (below referred to as Experiment 2), in which 12 annotators were given the task to annotate the same text about 350 words long. All annotators were familiar with the annotation workflow, even if to a different degree, and the research interest was to list points of disagreement of different kinds.

The total number of error spans marked in this text was 156. Of them, 57 were spotted by no more than 2 annotators, 23 were spotted by only 3 annotators, 30 errors were marked by at least 10 annotators of the 12 participants, and they all chose the same tag for these spans, and 6 areas spotted by at least 10 annotators were marked with different tags. What is left is 40 tags noticed by 4 to 9 annotators, and there are 19 among them in which the annotators agreed in their choice of tags (for the convenience of reference called in the graph "Part agree"). Fig. 6 shows the distribution of the spread of annotation decisions across the 12 annotators in the experiment.

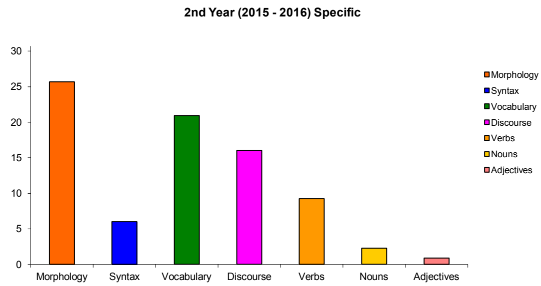

Fig. 5. Variation in the use of specific error tags in the current area of REALEC

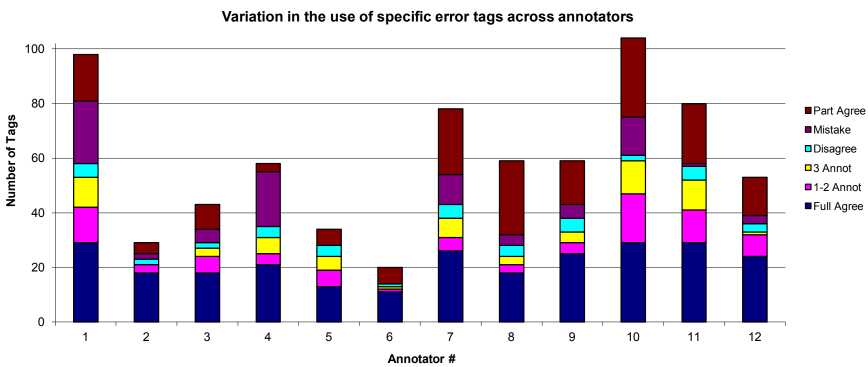

Fig. 6. Variation in the use of specific error tags by annotators in Experiment 2

To conclude, two annotation experiments demonstrated adequate reliability in the use of tags by REALEC annotators and in their approach to complicated errors. Hopefully, by increasing uniformity in annotation practice we will be able to approach automatisation of tagging in the learner corpus of student written works, and as a result, get closer to partially automated evaluation of student essays (as is indicated in McEnery, Tony and Richard Xiao (2011)). The corpus itself is a valuable pedagogical tool—for one, it provides a variety of possibilities for EFL instructors to create automated and semi-automated training exercises, as well as progress and placement tests on the basis of the mistakes annotated in learner texts in the corpus. The main feature of such exercises and tests is going to be their precision in targeting sharply at eliminating the specific mistakes that a particular group of learners is prone to making.

References

1. *Artstein R. & Poesio M.* (2008) Artstein, Ron and Massimo Poesio, Massimo Inter-coder Agreement for Computational Linguistics in Computational Linguistics 34(4), 2008, pp. 555–596.
2. *Bayerl P.* (2007) Bayerl, Petra Saskia Bayerl Identifying Sources of Disagreement: Generalizability Theory in Manual Annotation Studies in Computational Linguistics, 33(1), 2007, pp. 3–8.
3. *Braun S.* (2006) Braun, Sabine ELISA—a pedagogically enriched corpus for language learning Purposes in Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods Frankfurt/M: Lang, 2006, pp. 25–47
4. *Bryant Ch. & Hwee Tu Ng* (2015) Bryant, Christopher & Ng, Hwee Tou How Far are We from Fully Automatic High Quality Grammatical Error Correction?-in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 2015, pp. 697–707
5. *Chodorow, M. et al.* (2010) Chodorow, Martin, Gamon, Michael, Leacock, Claudia, & Tetreault, Joel Rethinking Grammatical Error Annotation and Evaluation with the Amazon Mechanical Turk—in Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 45–48
6. *Chodorow et al.* (2014) Chodorow, Martin, Gamon, Michael, Leacock, Claudia, & Tetreault, Joel. Automated Grammatical Error Detection for Language Learners 2014– in Series Title: Synthesis Lectures on Human Language Technologies Publisher: Morgan & Claypool Publishers, 2014
7. *Diez-Negrillo A. & Fernandez-Dominguez J.* (2006) Diez-Negrillo, Ana & Fernandez-Dominguez, Jesus Error-Tagging Systems for Learner Corpora—in Revista española de lingüística aplicada, # 19, pp. 83–102
8. *Glaznieks et al.* (2014) Glaznieks, Aivars, Nicolas, Lionel, Stemle, Egon, Abel, Andrea & Lyding Verena Establishing a Standardised Procedure for Building Learner Corpora in Journal of Applied Language Studies, vol. 8, 3, 2014, 5–20
9. *Granger S.* (2003) Granger, Sylviane 2003. The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research, TESOL Quarterly, 37, 3, p. 538–546
10. *Granger, Gilquin & Meunier* (2013) Granger, Sylviane, Gilquin, Gaëtanelle & Meunier, Fanny (eds) Twenty Years of Learner Corpus Research—Looking Back, Moving Ahead. Corpora and Language in Use—Proceedings 1, Louvain-la-Neuve, Presses Universitaires de Louvain, 2013
11. *Hovy E. & Lavid J.* (2010) Hovy, Eduard and Lavid, Julia Towards a “Science” of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics, in International Journal of Translation Studies 22(1), 2010: pp. 13–36.
12. *Hovy E.* (2015) Hovy, Eduard Corpus Annotation in Ruslan Mitkov (ed.) The Oxford Handbook of Computaional Linguistics Second Edition (2 ed.) Online publication November 2015

13. *Krippendorff K.* (2007) Krippendorff, Klaus Computing Krippendorff's Alpha Reliability available at http://web.asc.upenn.edu/usr/krippendorff/mwebreliability5.pdf
14. *Krippendorff K.* (2012) Krippendorff, Klaus. Content analysis: An introduction to its methodology. Sage, 2012.
15. *Kutuzov A., E. & O.* (2015) Kutuzov, Andrey, Kuzmenko, Elizaveta and Vinogradova, Olga Evaluating inter-rater reliability for hierarchical error annotation in learner corpora in the proceedings of 8th International Corpus Linguistics Conference, Lancaster 2015, pp. 211–214.
16. *Lee J. et al.* (2014) Lee, John, Yan Yeung, Chak, Zeldes, Amir, Reznicek Marc, Lüdeling Anke, and Webster, Jonathan CityU Corpus of Essay Drafts of English Language Learners: A Corpus of Textual Revision in Second Language Writing—electronic prepublication at https://corpling.uis.georgetown.edu/amir/pdf/annis_cityu_prepub.pdf
17. *McEnery, Tony and Richard Xiao* (2011) What corpora can offer in language teaching and learning. In Hinkel, E. (ed.), Handbook of Research in Second Language Teaching and Learning. London: Routledge.
18. *Poesio M., Bruneseaux F. & Romary L.*(1999) Poesio, Massimo, Bruneseaux, Florence & Romary, Laurent The MATE Meta-Scheme for Coreference in Dialogues in Multiple Languages In Proceedings of the ACL Workshop on Standards for Discourse Tagging, College Park, MD, pp. 65–74. Stroudsburg, PA: Association for Computational Linguistics.
19. *Pustejovsky J. & Stubbs A.* (2013) Pustejovsky, James & Stubbs, Amber (). Natural Language Annotation for Machine Learning. Sebastopol, CA: O'Reilly Media, 2013.
20. *Shtindlova B. et al.* (2014) Shtindlova, Barbara, Rosen, Alexandr, Hana, Jirka & Shkodova, Svatava—CzeSL—an error tagged corpus of Czech as a second language available at http://utkl.ff.cuni.cz/~rosen/public/2014-czesl-sgt-en.pdf
21. *Sorokin A. & Forsyth D.* (2008) Sorokin, Alexander, & Forsyth, David Utility Data Annotation with Amazon Mechanical Turk. In Proceedings of the First IEEE Workshop on Internet Vision at the Computer Vision and Pattern Recognition Conference (CPVR), 23–28 June Anchorage, AK, 1–8. Washington, DC: IEEE Computer Society, 2008.
22. *Stenetorp P. et al* (2012) Stenetorp, Pontus, Pyysalo, Sampo, Topić, Goran, Ohta, Tomoko, Ananiadou, and Tsujii, Jun-Ichi BRAT: A Web-Based Tool for NLP-Assisted Text Annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 102–107. Stroudsburg, PA: Association for Computational Linguistics.
23. *Yannakoudakis H.* (2013) Yannakoudakis, Helen Automated assessment of English-learner writing. Cambridge, 2013 https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-842.pdf
24. *Wilcock* (2009) Wilcock, Graham. Introduction to Linguistic Annotation and Text Analytics. San Rafael, CA: Morgan and Claypool Publishers, 2009.