Russian Learner Translator Corpus 2.0
https://rus-ltc.org

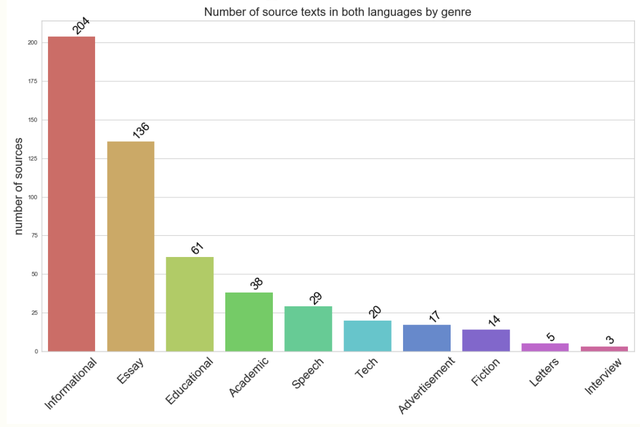
RusLTC

Parallel English<>Russian sentence-aligned corpus of multiple student translations.

Corpus size and metadata

- Basic statistics:
  - 2.3 mln words, 4.8K texts;
  - 402 English source texts, ≈ 8 Russian translations for each;
  - RU targets: ≈ 380 words.
- Metadata:
  - Translation grade;
  - Draft / final translation (editing effort);
  - Routine / Exam / Contest;
  - At class / at home.

- Source text genres:



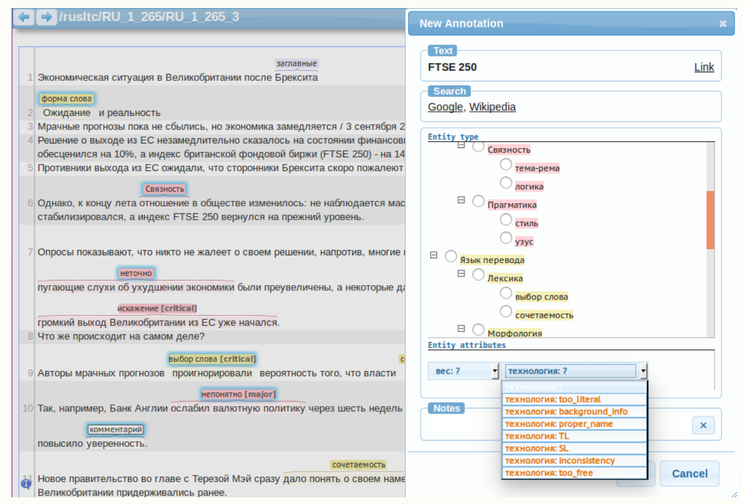
- Translators:
  - Gender-annotated;
  - 60% by advanced Translation Studies students;
  - 14 Russian universities.

Formats and availability

- Plain text documents;
- TMX (XML-based bitext format);
- stand-off error annotation files;
- fully public, downloadable, CC-BY-SA.

Error-annotated translations

- > 750 texts, 300K words, 16K error labels
- 30 error types, 3 weights, 10 tech attributes
- Annotated in **brat**:

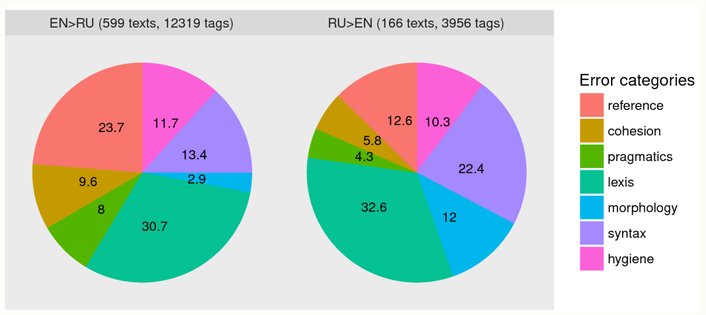


New search interface: ElasticSearch + Tsakonian Corpus Platform

- ***'Elected last October, he comes to office at a time when Latin America is in a state of upheaval.'***
  1. *'После выборов в октябре прошлого года он пришел к власти, в это время в Латинской Америке были беспорядки.'*
  2. *'Он был избран в октябре и сразу столкнулся с проблемой экономического спада в Латинской Америке.'*

- Search by lemmas and PoS:
  - *Do students find concise ways of rendering N+to+V_{inf}?*
  - *Do students overuse existential verbs?*
- Filtering by metadata (arbitrary subcorpora):
  - *Do advanced students use pronouns less?*
  - *Do they translate 'in fact' and 'people' in more different ways?*

Translation error types:



Some published research applications

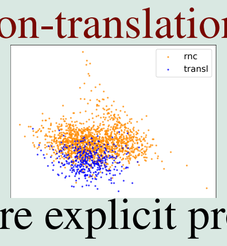
Testing distinctions

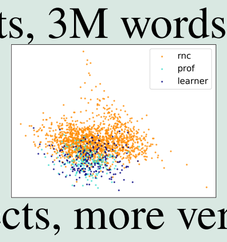
Learner vs. professional translations

- Sentence length, splitting.
- Types to tokens ratio (TTR).
- Frequent words distribution.
- Lexical density.
- Morphological forms distribution.
- Connectives and epistemic markers.

Translation types features

- learner: 200 texts, 206K words
- professional: 200 texts, 326K words
- non-translations: 1.5K texts, 3M words





More explicit pronoun subjects, more verb chains, more analytic passives, etc.

Textbooks vs. real errors

Data: (1) annotated multiple translations to difficult sentences from English mass-media, and (2) 38 textbooks for translation students (published in 1963-2016)

- Of the top 20 most active error-triggers only 3 are discussed in the textbooks!

Plans and Outlook

- Evaluate translation (alignment) accuracy: similarity between source sentences and their correct and erroneous translations.
- Lexical level: collocations.
- More syntactic suspects: word order and constructions.
- Comparison to machine translation:
  1. non-human semantic and coherence errors,
  2. professionalism is about fluency.

RusLTC Team
University of Tyumen, Russia

RUSSIAN FOUNDATION FOR BASIC RESEARCH
RFBR