Linguistic tendencies in English to Russian translation: the case of connectives

Maria Kunilovskaya

Dialogue, 23rd International Conference on Computational Linguistics and Intellectual Technologies

1 июня 2017 г.

Outline

Example 1 (unconventional representation)

Если раньше не так просто было получить заказ, то теперь получить за его выполнение деньги в разы сложнее.

Example 1 (unconventional representation)

Если раньше не так просто было получить заказ, то теперь получить за его выполнение деньги в разы сложнее.

Example 2 ("shining through"effect)

- Г-жа Мэй сказала, что переговоры не начнутся до 2017 года...
- В прошлом октябре
- Мир журналистики совершенно недооценен людьми сегодня.
- предстоит разобраться ... какие злодеи угрожают нам и какие суперспособности есть у них

Example 1 (unconventional representation)

Если раньше не так просто было получить заказ, то теперь получить за его выполнение деньги в разы сложнее.

Example 2 ("shining through"effect)

- Г-жа Мэй сказала, что переговоры не начнутся до 2017 года...
- В прошлом октябре
- Мир журналистики совершенно недооценен людьми сегодня.
- предстоит разобраться ... какие злодеи угрожают нам и какие суперспособности есть у них

Example 3 (defunct texture)

Букмекеры считают, что есть 40%-я вероятность того, что Великобритания не выйдет из состава ЕС до 2020 года.
Однако в скором времени сложилось впечатление, что экономика страны развивается лучше, чем предполагали.

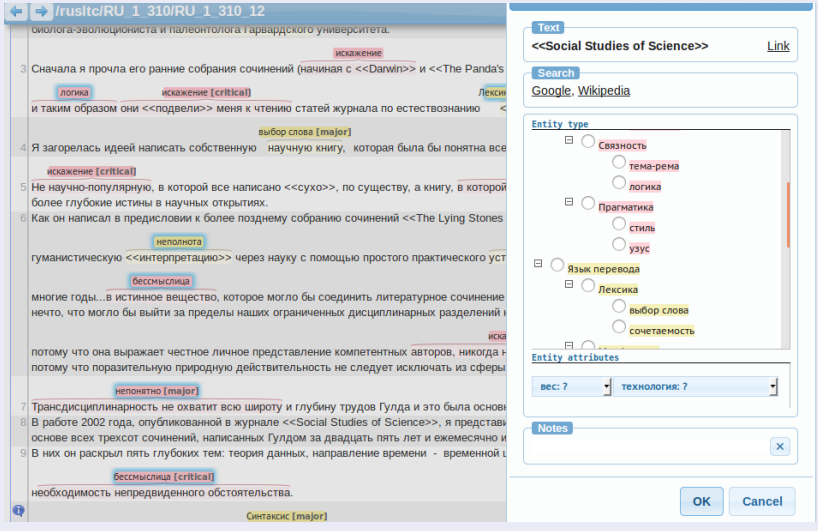
Translationese and translation quality

- Translationese is disfluency: translationese are linguistic features which distinguish translations from non-translations (Gellerstam, 1984)
- Fluency is quality: fluency is one of three major areas of TQA along with adequacy and faithfulness (Callison-Burch et al.,2007)
  - it can be assessed regardless of faithfulness/equivalence (Nord, 2003)
  - the most creative part of translation is producing the most standard representation for the given idea (Ryabtseva, 2013, p.45)
  - in mass media Idiom principle of text production is at its strongest (Alexeeva, 2004)
- Feature selection is based on translation universals theory
- Learner translations are a variety of translationese produced at the other level of competence (Previous work indicative of the task viability: Scarpa (2006), Kunilovskaya, Pariy (2016))

TQA in TS and CoLing/MT

Translation studies

holistic parametric evaluation and classification errors



Computational approaches

measure proximity of a candidate to a reference translation based on

- n-gram precision (BLEU/NIST/Meteor)
- word error-rate metric based on the Levensthein distance (PER/TERp)
- verb-frames and semantic roles comparison (HMEANT)
- post-editing effort (HTER) (Vela et al., 2014)

measuring translationese effect is similar to text quality assessment and translation detection tasks, although our goal is set within CoLing rather than NLP paradigm (Laippala et al., 2015)

Corpus-based translation studies and translation universals

Linguistic properties of translations as text in their own right

> *«features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems» (Baker, 1993 : 243)*

- describe and explain tendencies in translator's linguistic choices
- typify translations as a special mode of bi-lingual communication against non-translations in the target language
- revealed through corpus-based quantitative analysis

translationese (Gellestam, 1986),
third code (Frawley, 1984),
laws of translation (Toury, 1995)

Established translational tendencies

1. simplification
   lexical, syntactic, stylistic
2. explicitation
   spelling things out rather than leave them implicit (Blum-Kulka, 1986)
3. normalization and convergence/levelling-out
   «tendency to exaggerate features of the target language and to conform to its typical patterns» (Baker 1996: 183)
   «higher level of homogeneity of translations with regard to their own scores on given measures of universal features» (Laviosa 2002)
4. interference and transfer («shining through»)
   «in translation, phenomena pertaining to the make-up of the source text tend to be transferred to the target text» (Toury 1995: 275)
   «diverging frequencies of options existing in both languages are adapted in translated texts to those of the source language» (Teich 2003)

Linguistic indicators: Why connectives?

Frequency of connectives is an important textual property,

- known to differentiate registers within one language (Liu, 2008),
- reflecting text-structure conventions across languages (Fabricius-Hansen, 2005),
- effectively used in comparing (parallel) corpora (Cartoni, 2011),
- a traditional translationese indicator due to structural optionality (Olohan, 2001; Nakamura 2007; Kamenicka, 2007; Castagnoli, 2008; Denturk, 2009)

part of a project on features of learner translationese, inc. word order, collocation, sentence length

Connectives — lexical items, which

- form a functional subclass of discourse markers
- *«function like a two-place relation, one argument lying in the segment they introduce, the other lying in the prior discourse»* (Fraser, 1999)
- usually structurally optional, parenthetical
- do not add semantic content to the propositions, but make existing connection between them explicit
- facilitate text interpretation by overtly signaling its structure
- in grammar references and papers also referred to as *discourse markers, linking adverbials/sentence adverbials, particles, parenthetical expressions, prepositional phrases*

describe quantitative differences between learner and professional translations of mass-media texts from English into Russian as to the use of connectives and establish translational tendencies at work at different levels of competence

Major research questions:

1. Is there a cross-linguistic difference in the use of connectives in the reference corpora?
2. Do translations and non-translations feature differences with regard to the frequency distribution of individual connectives and their semantic groups? Translationally distinctive connectives? How can patterns of overuse and underuse be explained from cross-linguistic perspective (interference, convergence)?
3. What is the relation between the two types of translation into Russian (learners and professionals) in terms of explicit text cohesion (as compared to sources and reference corpus)?

Corpus resources

Test corpora	No. of texts	Size[1]
Learner corpus		
sources	208	214,237
translations	208	197,646
Professional corpus		
sources	200	359,255
translations	200	330,973

Reference corpora	No. of texts	Size
RNC selection	1,562	3,129,731
BNC selection	–	5,735,974

All samples are made comparable in genre at building stage.

Russian Learner Translator Corpus

[1] after preprocessing and linguistic annotation

Search lists

go to Appendix 1

Two independent search lists (+semantic groups) based on

English (105 items)

Biber et al. (1999), Fraser (1999, 2006), Liu (2008), Meyer and Webber (2013), Swan (1992)

Russian (95 items)

Berson et al (1984), Kogut (2014), Novikova (2008), Priyatkina (2007, 2015), Russian Grammar (Shvedova 1980)

Selection criteria: Items included

- have limited variation, mostly deictic (e.g. for this|that reason)
- allow positional or punctuational disambiguation with homonyms
- are structurally optional (usually parenthetic)
- have relational semantics reflected in dictionaries (Oxford, Macmillian; Burtseva, 2010)

Extraction procedure

Issues with matching substrings method of extraction from raw text

- *а именно* ← *Выбор членами отдела именно этого документа*; *in addition* ← *to gain additional skills*
- *кроме того* and *кроме этого* counted separately

Redefine default POS annotations and get tokenized and lemmatized MWE for search items

- check for possible variation and combinations (*всё-таки* and *все-таки*; *а поэтому*, *а главное*)
- use positional and punctuational features to disambiguate (*by the way* ← *his love for his wife was deepened by the way she stood by him during his years in prison*
- account for variation in default POS tags (*во-первых* is tagged R and Afpmpgf; *all* – as DT, PDT and RB

Which frequencies?

standard measure

connectives per sentence (ips) x 100

freq. of all items by document

degree of explicit cohesion

freq. of items by semantic group

cross-linguistic differences in the preferred type of cohesion

individual frequencies of connectives, frequency bands

overused and underused items

verification

analysis of parallel data

Semantic groups of connectives

dealing with polyfunctionality (Mayer and Webber 2013);

draws on taxonomies by Fraser (2006), Halliday and Hasan (1976), Biber (1999), in Russian: Berson (1984); Inkova-Manzotti (2001)

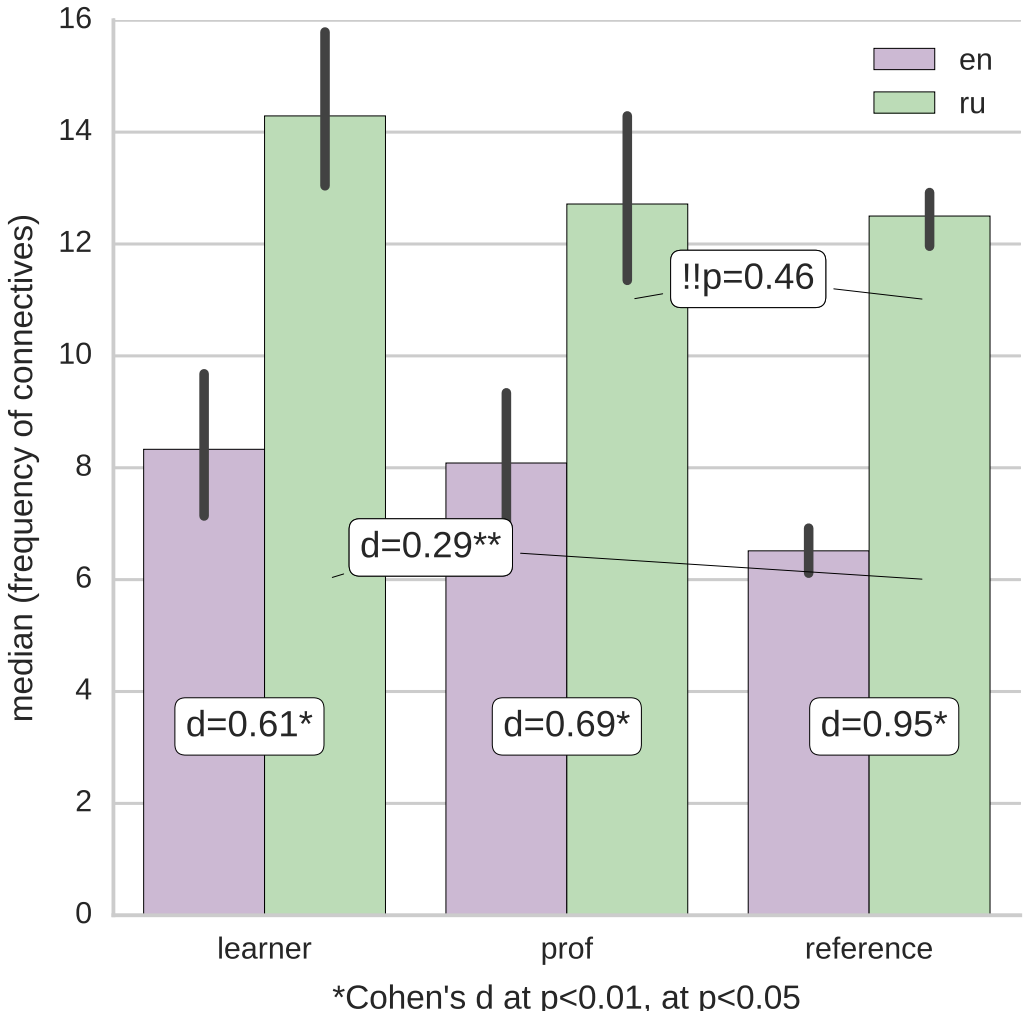
Semantics	Subtype	Eng (105)	Rus (95)
elaboration	reformulation extension focusing	35	41
inference	cause/reason/result conclusion, deduction	18	15
contrast	contrast concession	21	16
sequential	order and listing change of topic summary/generalising	30	24

Basic descriptive statistics

corpus	raw freq	median(nfreq)	SD(ips)	Cohen's D[2]
Learner corpus				
sources	976	8.33	9.0	0.61
translations	1,494	14.29	11.6	
Professional corpus				
sources	1,306	8.08	6.22	0.69
translations	2,115	12.72	7.93	
Reference corpora				
BNC	12,439	6.52	3.13	0.95
RNC	22,821	12.50	7.78	

[2]significant at $p<0.01$, measured by (un)paired two-tail Wilcoxon rank test

Medians for average conn. frequencies in translations, sources and reference

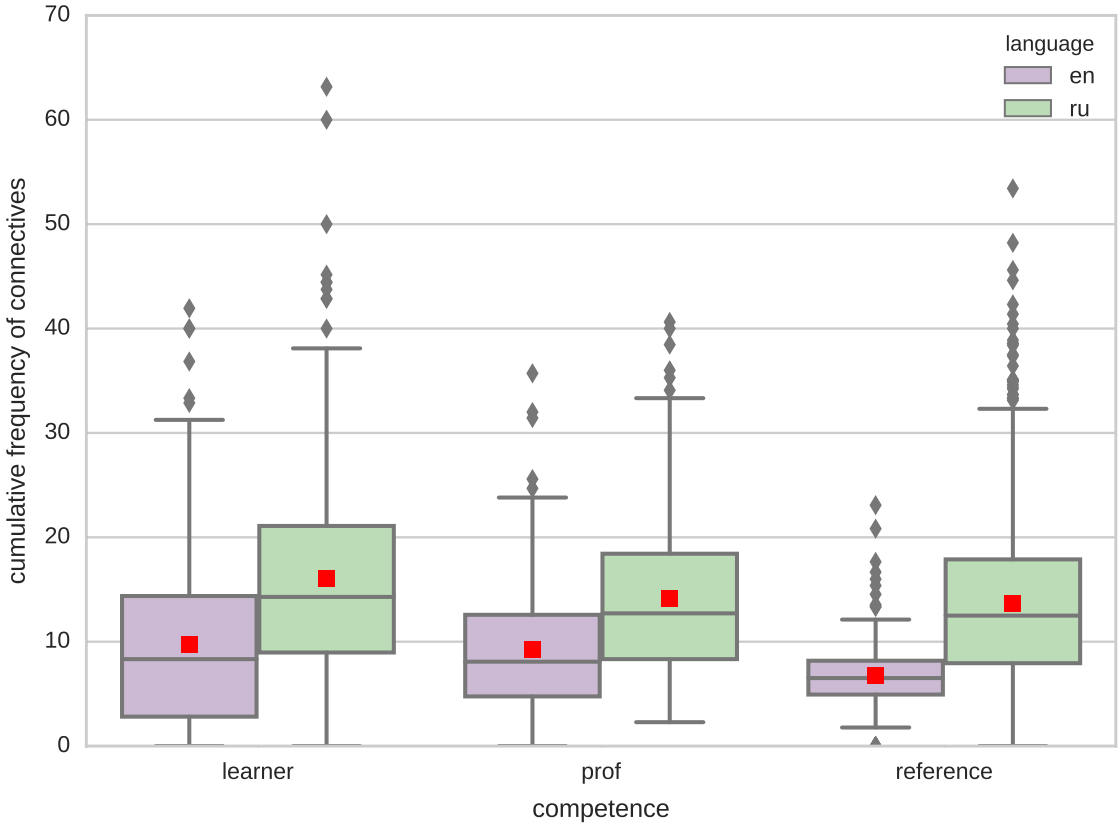


Freq. of connectives in translations against reference *(normalized over no. of sentences per text x 100)*

Wilcoxon rank test

learner vs. RNC
p = 0.01343
Cohen's D = 0.29
prof vs. RNC
p = 0.3409
sources vs. BNC
p = 5.946e-05
Cohen's D = 0.4977

- no standardization
- mild overuse by learners: normalization? explicitation?



overused items[3]

learners	professionals
например	вместо этого
вместо этого	*также*
при этом	на самом деле
на самом деле	однако
однако	*в конце концов*
затем	
6	5

at $p < 0.05$ (Wilcoxon rank sum test on normalized frequencies of 95 items across texts in the corpora, in the ascending order of p-value; compare to LL results)

overused are from 3d freq quartile; underused items come from the 1st freq quartile

no 'unique translationese' (Cartoni, 2011)

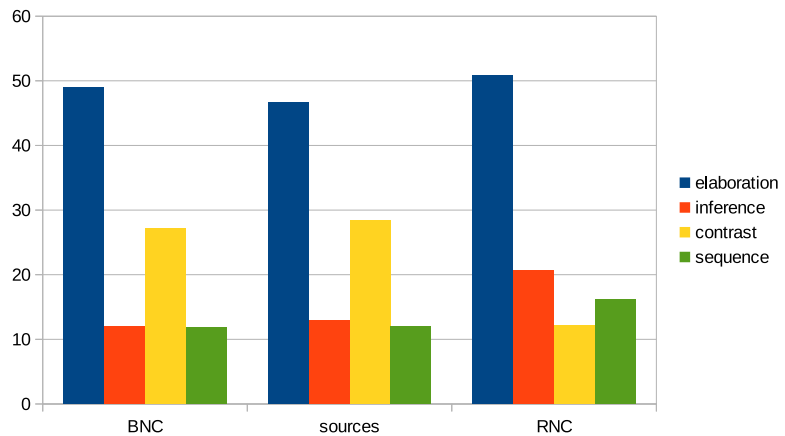
underused items

learners	professionals
поэтому	так
причем	прежде всего
так	кроме этого
то есть	кстати
впрочем	причем
ведь	например
кстати	то есть
кроме этого	в частности
прежде всего	ведь
в частности	в связи с этим
наконец	впрочем
значит	
вместе с тем	
в связи с этим	
во-первых	
15	11

[c] b - elaboration; y - contrast; g - sequence, r - inference

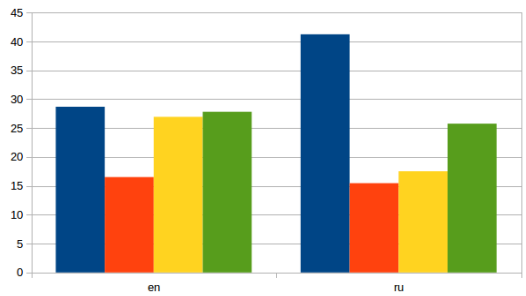
Preferred text relations to be signaled

Do translations use the same types of connectives as sources or reference? ⇐
Are there differences between languages in the semantic type of connectives used?



Ratio of semantic types of connectives in the reference corpora (based on av. rel. freq)

not immediately related to the size of search lists



Ratio of list sizes for sem.groups

This can predict overuse of contrastiv markers under the hypothesis of interference and inferential markers in case of normalization

Table 4. Aggregated frequencies for semantic groups of connectives across the corpora

Semantic type	No. of items		leST	learners	proST	pro	BNC	RNC
	EN	RU						
Elaboration	35	41	550	756	759	1,037	8,129	11,906
			56.4%	50.6%▼	56.4%	49.0%▼	65.4%	52.2%
Inference	18	15	124	221	132	313	820	4,780
			12.7%	14.8%▲↓	12.7%	14.8%▲↓	6.6%	20.9%
Contrast	21	15	198	271	238	399	2495	2,400
			20.3%	18.1%▼↑	20.3%	18.9%▼↑	20.1%	10.5%
Sequential	30	24	104	246	177	366	993	3,735
			10.7%	16.5%▲≈	10.7%	17.3%▲≈	8.0%	16.4%
Total (100%)	105	95	976	1,494	1,306	2,115	12,437	22,821

▼ ▲ change in frequency distribution compared to sources

↓ ↑ change in frequency distribution compared to reference

(1) Cross-linguistically

1. English is «less explicitly cohesive» (wrt the inspected type of cohesion) → *A problem in ST comprehension and requires adaptation to TL conventions*

Example 4

Events in Brazil have revived the market in economic gloom. Newspapers are again warning of global recession, slump, or even depression. But when does a recession become a depression? (Economist 1999)

2. Relative difference in preferred type of relations marked (contrast in English, inference and sequence in Russian) → *conditions for adaptation shifts*

(2) tendencies in EN>RU translationese

1. normalization: increase in overall frequencies

Example 5

Anyone who, as a child, possessed a Junior Conjuror's Set will have learned two simple lessons about magic. Audiences long to be deceived and are invariably disappointed when they are told how the trick is done.

Каждый у кого был набор юного фокусника в конечном итоге поняли два основных правила магии - не бойтесь обманывать зрителя, ведь они ожидают увидеть чудо, и ни в коем случае не рассказывайте секрет фокуса - их это сильно разочарует.

(2) tendencies in EN>RU translationese

1. normalization: increase in overall frequencies

Example 5

The Nobels are a great way to get people interested in science, they'll say, and it's good that we have them.	Говорят, что Нобелевская премия действительно побуждает людей интересоваться наукой, и поэтому здорово, что она есть.

(2) tendencies in EN>RU translationese

1. normalization: increase in overall frequencies
2. normalization: cross-linguistic shift towards adding markers of inference and sequence (as in Examples 5) + decrease in contrastive markers

(2) tendencies in EN>RU translationese

1. normalization: increase in overall frequencies
2. normalization: cross-linguistic shift towards adding markers of inference and sequence (as in Examples 5) + decrease in contrastive markers
3. interference: overuse of lower-freq., underuse of higher-freq. items (ex. *впрочем*)

(2) tendencies in EN>RU translationese

1. normalization: increase in overall frequencies
2. normalization: cross-linguistic shift towards adding markers of inference and sequence (as in Examples 5) + decrease in contrastive markers
3. interference: overuse of lower-freq., underuse of higher-freq. items (ex. *впрочем*)
4. interference: surplus of marked contrast typical for English (*однако* is triggered by sentence-initial *but* (42%) and *however* (36%) + *nevertheless, yet, instead, though*) and lack of inferential connectives (all of them underused, except *в результате* all triggered by a variety of expressions)

(2) tendencies in EN>RU translationese

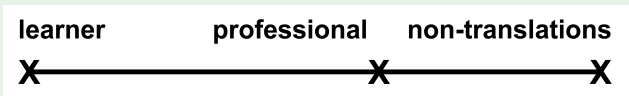
1. normalization: increase in overall frequencies
2. normalization: cross-linguistic shift towards adding markers of inference and sequence (as in Examples 5) + decrease in contrastive markers
3. interference: overuse of lower-freq., underuse of higher-freq. items (ex. *впрочем*)
4. interference: surplus of marked contrast typical for English (*однако* is triggered by sentence-initial *but* (42%) and *however* (36%) + *nevertheless, yet, instead, though*) and lack of inferential connectives (all of them underused, except *в результате* all triggered by a variety of expressions)

translations are in the gap btw two languages

(3) learners vs. professionals

- same tendencies, which are inevitably stronger at lower level of competence

Example 5 (также is annoying in learner translations)

It was twice more frequent in learner translations, with professional translations counts in between 

- no significant differences, no specific tendencies
- other and non-list markers of discourse converge to one connective

Example 6

The scene can be interpreted otherwise	Однако эту сцену можно понять по-другому.

- a mild trend towards explicitation and simplification

Linguistic tendencies in English to Russian translation: the case of connectives

Maria Kunilovskaya

Dialogue, 23rd International Conference on Computational Linguistics and Intellectual Technologies

1 июня 2017 г.