

Лингвистические свойства переводов методами корпусной лингвистики: *год спустя*

Мария Куниловская

Индустрия перевода, ПНИПУ

Пермь, 06 июня 2018

Содержание

- 1 Языковая специфика переводов как основание оценки
 - Мотивация и цель
 - Теория и методология
- 2 Синтаксические признаки студенческих переводов
 - Research design
 - Корпусные данные и пространство признаков
 - Классификация
 - Результаты
 - Case studies
- 3 Другие результаты
- 4 References

Можно ли формализовать оценку качества перевода?

Область исследований:

- эмпирическое изучение переводов, корпусная и компьютерная лингвистика

Цель проекта:

- построить модель признаков переводного текста для аппроксимации его экспертной оценки
- верифицировать описанные в литературе отклонения переводов от ожидаемой нормы и квантифицировать их

Исходные посылки:

- 1 эксперт отличает хороший перевод от плохого по объективным формализуемым признакам
- 2 текст перевода можно оценить без соотнесения с оригиналом
- 3 профессиональные переводы лучше студенческих

Практический результат:

- автоматическое ранжирование студенческих (конкурсных) переводов по интерпретируемым признакам, согласующееся с экспертной оценкой

Теория (1): основные критерии качества

- ❶ адекватность (fitness for communicative purpose)
[House, 2001, Nord, 2006]
- ❷ точность (fidelity, semantic accuracy)
- ❸ **соответствие языковой норме и узусу (readability)**
 - студенческие переводы отличаются от профессиональных именно по степени читабельности (Carl and Buch-Kromann, 2010)
 - субъективность и нечеткость разграничения этих критериев [Callison-Burch et al., 2007]
 - возможность формализации отклонений от нормы (“переводизмов”)
 - норма ожидания заказчика/потребителя (Крылов (04.06.2018): “трудно воспринимать”)

Теория (1, cont.): подходы к оценке

- **академические** (Латышев 2003, Княжева 2015) и индустриальные (BLEU, TransCheck, Multidoc, LISA QA)
- **дескриптивные** и прескриптивные
- **продукт** и процесс (ISO 17100:2015)

Переводоведение

- “холистические” параметрические системы (напр., DipTrans, [Williams, 2009])
- классификации ошибок (напр., ATA Framework, MeLLANGE, Бузаджи 2009)

НО: субъективизм,
трудозатраты, неформализуемо

Компьютерная лингвистика

измерение лингв. расстояния

- (в МП) до референтного перевода (BLEU, Meteor see [Vela et al., 2014])
- до сравнимых непереводных текстов на ПЯ (норма ожидания, DTS [Toury, 1995], многомерный анализ [Biber, 1988])

Теория (2): корпусное переводоведение и переводизмы

Дескриптивизм: изучение особенностей языка переводов как подсистемы ПЯ

translationese [Gellerstam, 1986],

third code [Frawley, 1984],

laws of translation [Toury, 1995]

“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]

Тенденции, влияющие на языковой выбор переводчика:

- симплификация
- эксплициация
- интерференция
- нормализация
- гипотеза уникальных единиц

Количественные исследования переводов: направления

- соотнесение закономерностей процесса и результата
[Carl and Buch-Kromann, 2010]
- изучение тенденций в переводе
[Laviosa, 1998, Mauranen, 2004, Xiao et al., 2006, Bernardini, 2007, Hansen-Schirra, 2011]
- выявление переводов (translationese detection:
[Baroni and Bernardini, 2006, Ilisei, 2012, Popescu, 2011, Volansky et al., 2015])
- изучение вариативности языка переводов
 - жанровой [Kruger and van Rooy, 2010, Neumann, 2013],
 - стилометрия [Rybicki, 2005],
 - в зависимости от направления [Evert and Neumann, 2017],
 - **квалификации переводчика**
[Rabadán et al., 2009, Rubino et al., 2016, Sutter et al., 2017]

Общая методика эмпирических исследований

- 1 формулировка верифицируемой гипотезы и общий дизайн эксперимента
- 2 выбор или создание рабочих корпусов (коллекции текстов в машиночитаемом формате)
- 3 выработка способов представления данных (разметка?) и извлечения лингвистической информации
- 4 статистический и качественный анализ
- 5 интерпретация полученных результатов с целью их переработки в знание

Общие вопросы-1

Известно, что машина лучше чем человек отличает переводы от не-переводов [Baroni and Bernardini, 2006], используя при этом малоинформативные признаки (сочетания символов: 'дж', 'ейт').

- Сможет ли она различать хорошие и плохие переводы, если ей показать, на что смотреть?
- Можно ли обучить классификатор на интерпретируемых признаках (например, *синтаксические*) и извлечь из эксперимента знание о синтаксических особенностях переводов?

Общие вопросы-2

- Как представить текст в виде статистики синтаксических отношений?
Грамматика зависимостей и приложения для автоматического синтаксического анализа текста из проекта Universal Dependencies
- Где взять корпуса переводов разного качества?
Представим качество как функцию от профессиональной компетенции (студенты vs. профессионалы), т.е. зададим классы сравниваемых текстов экстралингвистически!

Why syntax?

- лексические ошибки часто носят локальный характер, **некоммуникативный порядок слов, отсутствие когезии? неграмматичность препятствуют пониманию смысла текста;**
- предполагают, что интерференция на синтаксическом уровне чаще остается незамеченной переводчиком [Mauranen, 2004];
- чем профессиональные переводчики проводят более глубокие синтаксические преобразования (Munday, 2013);
- синтаксические явления частотнее и более независимы от темы текста.

Обобщаем

Итак, задачи

- 1 построить алгоритм машинного обучения, который, принимая на вход синтаксическую статистику, сможет с приемлемой точностью классифицировать переводы/не-переводы и студентов/профессионалов
- 2 if successful, определить наиболее значимые для классификатора синтаксические признаки и провести их качественный анализ

Данные и признаки

Russian Learner Translator Corpus

Description

- > 2,2 млн слов
- параллельный множественный корпус, выровненный на уровне предложений
- доступен для скачивания и через интерфейс поиска <https://rus-ltc.org>
- EN < > RU
- 10 видов метаданных
- 11 жанров
- переводы студентов ПиП из 15 вузов России
- около 700 текстов размечено по ошибкам



Нам нужны студенческие переводы!

Данные и признаки

Корпусные ресурсы (EN > RU)

Тестовые	Текстов	Слов
Учебный корпус (RusLTC)		
Оригиналы	200	223K
Переводы	200	205K
Профессиональный корпус		
Оригиналы	200	343K
Переводы	200	321K



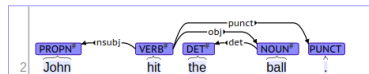
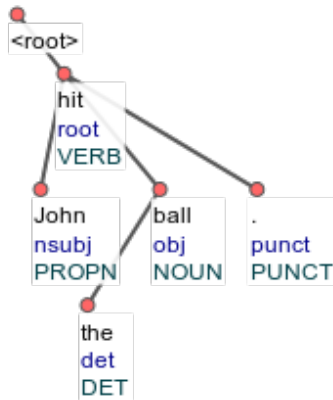
Референтный	Текстов	Слов
НКРЯ (не-переводы)	1,5K	3M

- все тексты одного типа: информативный газетный текст
- синтаксический анализ *UDPipe 1.2*
[Straka and Straková, 2017], русс. модель обучена на *UD 2.1 SynTagRus* treebank [Droganova and Zeman, 2016].

Пространство признаков

3 группы, 45 признаков

- 34 синтаксических отношения, (напр., дополнение, обстоятельство, предикат), представленных как средние по тексту вероятности их появления в предложении
- 7 графовых признаков, отражающих структурную сложность предложения
- 4 меры линейной и иерархической сложности дерева



Данные и признаки

формат данных: CONLL-U format

Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel	Deps	Misc
# newdoc									
# newpar									
# sent_id = 1									
# text = John hit the ball.									
1	John	John	PROPN	NNP	Number=Sing	2	nsubj	_	_
2	hit	hit	VERB	VBD	Mood=Ind Tense=Past VerbForm=Fin	0	root	_	_
3	the	the	DET	DT	Definite=Def PronType=Art	4	det	_	_
4	ball	ball	NOUN	NN	Number=Sing	2	obj	_	SpaceA
5	.	.	PUNCT	.	_	2	punct	_	SpaceA

Universal Dependency Relations

- 17 universal PoS tags + morph.features
- 43 and 39 relations for English and Russian

UDpipe - это программа (алгоритм) для токенизации, частеречной разметки и синтаксического парсинга текста. Для обучения алгоритма используются уже размеченные корпуса = **трибанки (treebanks)**.

A402

f_x ∑ =

450539

	A	B	C	D	E	F	G	H	I
1	doc	group	acl	acl:relcl	advcl	advmod	amod	appos	aux
2	RU_1_57_1	learner	0.0087968117	0.014852673	0.015991046	0.0363778827	0.1527787896	0.0026455026	0.00226757
3	RU_1_45_1	learner	0.0138441341	0.0502413847	0.0259468045	0.0658873853	0.1219661421	0.0023809524	0.00680272
4	RU_1_88_1	learner	0.0176776927	0.0118843425	0.0152754528	0.0435337212	0.1582716811	0.0032194998	0.00124601
5	RU_1_76_1	learner	0.0169692989	0.0094170535	0.007060689	0.0339626759	0.095112389	0.0164644179	0.00413223
6	RU_1_147_1	learner	0.0153122825	0.0193451272	0.0192741938	0.0406851866	0.1375439641	0.0043619162	0.00141043
7	RU_6_2_1	learner	0.0129451567	0.0016025641	0.0361581668	0.0509277527	0.1088812716	0.0236111111	
8	RU_1_47_1	learner	0.0160584692	0.0046296296	0.0206540574	0.0312393287	0.117428772	0.0204530088	
9	RU_1_175_1	learner	0.013761066	0.0054238571	0.0111875091	0.1016050359	0.1138935637	0.0016447368	
10	RU_1_244_1	learner	0.0356408867	0.0343120436	0.0040257649	0.0697816139	0.0977994589	0.0140786749	
11	RU_1_133_1	learner	0.0057142857	0.0156753247	0.0122451112	0.045948993	0.1345763302	0.0128078818	
12	RU_1_305_1	learner	0.0147880891	0.0184500643	0.0271327271	0.0458695468	0.1392582305	0.0032608696	0.00417732
13	RU_4_36_1	learner	0.0206223853	0.0064624963	0.0281750668	0.0771806365	0.1279965125	0	0.01712393
14	RU_1_3_1	learner	0.0125732111	0.0059202059	0.0310243016	0.0835850667	0.0741324058	0.0018018018	0.01103983
15	RU_1_89_1	learner	0.0186982636	0.0104424275	0.0099978967	0.0349461487	0.1408293295	0.0035593365	0.00229036
16	RU_1_245_1	learner	0.0179700368	0.0047775688	0.0069964097	0.0367428234	0.1160052658	0.0056722689	0.00544217
17	RU_1_82_1	learner	0.0187711112	0.0097804785	0.009375	0.0519264685	0.1297282458	0.0104166667	
18	RU_1_283_1	learner	0.011773668	0.01090179	0.00959896	0.032136764	0.0838342809	0.0106936011	0.00227574
19	RU_1_172_1	learner	0.0097622863	0.0111133513	0.0219373363	0.0545044321	0.1153662636	0.0216292435	
20	RU_3_10_1	learner	0.004743693	0.0043650794	0.0311601392	0.0433631178	0.1068938904	0.0081012717	
21	RU_1_75_1	learner	0.0062111801	0.0119047619	0.0264529725	0.0356134549	0.1407585686	0.0062111801	
22	RU_4_45_1	learner	0.0088460144	0.0036263414	0.0078453231	0.0455405115	0.1339718437	0	
23	RU_1_7_1	learner	0.0209222948	0.0164390385	0.0154118485	0.0467225866	0.0903949856	0.0208025169	
24	RU_5_6_1	learner	0.0336626984	0.0142857143	0.0196626984	0.0485833333	0.1231904762	0.0057142857	
25	RU_1_144_1	learner	0.0180305571	0.0276736097	0.012244898	0.0391945356	0.1730490532	0	
26	RU_1_286_1	learner	0.0214891359	0.0200315903	0.0151487738	0.055452578	0.1165880309	0.0089074404	0.0043648
27	RU_4_16_1	learner	0.0131147661	0.0071151924	0.0202243941	0.0416895305	0.1491130459	0.0048748353	0.00255754
28	RU_1_48_1	learner	0.0171713892	0.0161550571	0.0164787485	0.0231847698	0.1155864926	0.0013774105	0.00090909
29	RU_1_182_1	learner	0.0092710363	0.0159329036	0.0068700636	0.0795678011	0.1417149977	0.0021200416	0.00393516
30	RU_1_92_1	learner	0.0146489665	0.0087114571	0.0100406684	0.0390717807	0.1072315396	0.0093852797	0.00153822
31	RU_1_260_1	learner	0.0050096962	0.0130954516	0.0204872742	0.0849611864	0.0909587718	0.0016722408	
32	RU_4_33_1	learner	0.0156999331	0.0029429861	0.0067077029	0.0694142569	0.1367623525	0.006802935	0.00154798
33	RU_1_212_1	learner	0.0088461538	0.0101111111	0.0204711538	0.0639308877	0.0643317577	0.0033333333	
34	RU_1_243_1	learner	0.0151858389	0.0061724655	0.0069244161	0.0236089218	0.1245891067	0.0047969153	0.00051282
35	RU_1_288_1	learner	0.0170452586	0.0070495399	0.01439263	0.0562295512	0.1040662467	0.0116076464	0.01304347
36	RU_1_19_1	learner	0.0115948306	0.0060691436	0.006897571	0.0589853964	0.1420958539	0	0.00274725
37	RU_1_207_1	learner	0.0143935349	0	0.0104026408	0.1145634133	0.0868461604	0	

Обучение с учителем, задача классификации

- ядро классификатора - **метод опорных векторов** (SVM) со взвешенными коэффициентами для классов
- мера точности классификации: macro F1 на кросс-валидации ($k=10$)
- сравнение с альтернативными подходами (baseline):
 - результаты того же классификатора, обученного на сочетаниях частеречных тегов [Pastor et al., 2008]
ex. '*SCONJ PROPN VERB*', '*NOUN NOUN ADJ*'.
- способ определения наиболее информативных признаков:
 - дисперсионный анализ **корреляции признака с меткой класса** (ANOVA F-value)

Результаты

Результаты 5 экспериментов по сравнению с альтернативой

Macro-F1 for different feature sets and classification setups

Binary classification					3-class
	<i>transl/RNC</i>	<i>learner/RNC</i>	<i>prof/RNC</i>	<i>learner/prof</i>	
10 best features					
PoS baseline	0.735	0.738	0.658	0.791	0.603
Syntax	0.818	0.796	0.740	0.721	0.635
all features					
PoS baseline	0.820	0.820	0.797	0.806	0.707
Syntax	0.866	0.841	0.871	0.703	0.707

Интерпретация: классификация на синтаксических отношениях

- не хуже, результаты сравнимы с используемыми подходами
- требует в разы меньше вычислительных мощностей

Case studies

Обобщение результатов статистического анализа и результатов экспериментов

Анализ проводился только для значимых признаков, общих для классификаций **переводы vs. не-переводы** и **студенты vs. профессионалы**.

- Во всех экспериментах в топ-10 одни и те же синтаксические отношения;
- Большинство из них имеют большую вероятность в переводах;
- Задача различения студенческих и профессиональных переводов - самая сложная
- Обнаруженные закономерности:
 - **повышенная сложность синтаксиса** = большее число зависимых клауз и сложных глагольных предикатов:
advcl, acl:relcl, ccomp, xcomp.
 - **использование пассивов** = вероятность *aux:pass, nsubj:pass* выше.

Синтаксическая усложненность (1)

В переводах (особенно студенческих)

- больше придаточных: чаще используются **advcl**, **acl:relc**;
- чаще эксплицитно выражены субъекты (особенно прономинализированные) **nsubj**:

Example (more of **nsubj**)

... человек на улице не думает о ЕС, **когда он** входит в торговый центр.

... the man in street not think about EU, **when he** enters in shopping center.

(Source: ... the man on the street is not thinking about EU **as he** enters a shopping centre.)

- чаще эксплицитно выражены подчинительные отношения **mark**.

Синтаксическая усложненность (2)

Переводы выдает перенос **xcomp** - сложного глагольного/модального сказуемого

Example (модальные, каузативные, аспектальные глаголы)

- ... **пытается** заставить их сесть ...
... **try to make** them sit down ...
- ... **может заставить** потерять управление.
... **can cause** to lose control.

Интерференция

- **EN > RU** (x 1.4)
- **Translations > non-translations** (1.5 times)

Симплификация

Меньшее лексическое разнообразие:

- top 25 heads cover 83% (learners), 77% (prof), 73% (RNC).

Форма пассива

Переводы vs. не-переводы

- **Translations use less passives in general:**
down to 15% (learners) and 11% (prof) to the number of sentences from 18.6% in RNC.
- **But more of these passives are analytical:**
up to 38% (learners) and 35% (prof) from 25% in the RNC.

Example (analytical passives)

- ... аргумент может **быть учтён**...
... this argument can **be considered** ...
- эти недостатки **были компенсированы** за счет...
... these shortcomings **were remedied** by ...
- **Были проведены** научные исследования, целью которых ...
Was conducted research to find out ...

What we have learned

- 1 Синтаксическое представление переводов позволяет отразить отличия переводов и не-переводов
Наш результат $F1 = 86.6\%$ сравним с
 - Bag-of-words 87.19% [Rubino et al., 2016];
 - PoS+feat+syntax 86.26% [Laippala et al., 2015];
 - PoS trigrams on our data 82% after [Pastor et al., 2008].
- 2 В переводах с английского в предложениях **больше клауз**, чем в непереводных русских текстах
- 3 В переводах **чаще встречаются полностью реализованные структуры** (без эллипса), предложения с составным сказуемым и аналитическими пассивами.

Студенты vs. проф. vs. непереводы (корпусные методы)

Статистические различия по параметрам:

- ### 1 Длина предложения

learner professional non-translations

X ————— X ————— X

- 2 Плотность информации (соотношение знаменательных и служебных слов)

learner professional non-translations

X X X

- ### 3 Лексическое разнообразие

learner professional non-translations

X-----X-----X

- соотношение словоформ и словоупотреблений (TTR на леммах)
- соотношение высокочастотных и низкочастотных слов

- #### 4 Частотности коннекторов и некоторых частей речи

Спасибо,
что подумали над моими вопросами:

- формализуемо ли языковое качество переводов?
- коррелируют ли статистически значимые различия с воспринимаемым/предполагаемым качеством?

Вопросы?

Мария Куниловская и Ко
mkunilovskaya@gmail.com

Russian Learner Translator Corpus

References IV

- ▶ Hansen-Schirra, S. (2011). Between normalization and shining-through. Specific properties of English-German translations and their influence on the target language. *Multilingual Discourse Production: Diachronic and . . .*, pages 133–162.
- ▶ House, J. (2001). Translation Quality Assessment: Linguistic Description versus Social Evaluation. *Meta: Journal des traducteurs*, 46(2):243.
- ▶ Ilisei, I.-n. (2012). *A machine learning approach to the identification of translational language: an inquiry into translationese*. Doctoral thesis, University of Wolverhampton.

References V

- ▶ Kruger, H. and van Rooy, B. (2010). The features of non-literary translated language: a pilot study. *The International Symposium on Using Corpora in Contrastive and Translation Studies*, (2001):1–29.
- ▶ Laippala, V., Kanerva, J., Missilä, A., Pyysalo, S., Salakoski, T., and Ginter, F. (2015). Towards the classification of the Finnish Internet Parsebank: Detecting translations and informality. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 107–116. Linköping University Electronic Press, Sweden.
- ▶ Laviosa, S. (1998). Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. *Meta: Journal des traducteurs*, 43(4):557.

References VI

- ▶ Mauranen, A. (2004). Corpora, universals and interference. In *Translation universals: Do they exist*, pages 65–82. J. Benjamins. Amsterdam.
- ▶ Neumann, S. (2013). LSB2013 conference - Genre- and Register-related Text and Discourse Features in Multilingual Corpora 11-12 January 2013 - Institut libre Marie Haps, Brussels (Belgium) - www.mariehaps.be/lb2013. (January):11–13.
- ▶ Nord, C. (2006). Translating as a purposeful activity: a prospective approach. *TEFLIN Journal: A publication on the teaching and ...*, 17(2):131–143.

References VIII

- ▶ Rubino, R., Lapshinova-Koltunski, E., and Van Genabith, J. (2016). Information Density and Quality Estimation Features as Translationese Indicators for Human Translation Classification. In *HLT-NAACL*, pages 960–970.
- ▶ Rybicki, J. (2005). Burrowing into translation: Character idiolects in henryk sienkiewicz’s trilogy and its two english translations. *Literary and Linguistic Computing*, 21(1):91–103.
- ▶ Straka, M. and Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.

References IX

- ▶ Sutter, G. D., Cappelle, B., and Looock, R. (2017). Towards a corpus-based, statistical approach to translation quality: Measuring and visualizing linguistic deviance in student translation. *Linguistica Antverpiensia, New Series*, pages 25–39.
- ▶ Toury, G. (1995). *Descriptive Translation Studies - and Beyond*. John Benjamins Publishing Company.
- ▶ Vela, M., Schumann, A.-C., and Wurm, A. (2014). Beyond Linguistic Equivalence. An Empirical Study of Translation Evaluation in a Translation Learner Corpus. In *Proceedings of the EACL Workshop on Humans and Computer-assisted Translation*, pages 47 – 56.

References X

- ▶ Volansky, V., Ordan, N., and Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities*. 30(1):98–118.
- ▶ Williams, M. (2009). Translation quality assessment. *Mutatis mutandis*, 2(1):3–23.
- ▶ Xiao, R., He, L., and Ming, Y. (2006). In Pursuit of the Third Code: Using the ZJU Corpus of Translational Chinese in Translation Studies. In Xiao, R., editor, *Using Corpora in Contrastive and Translation Studies*, pages 182–214. Cambridge Scholars, Newcastle.