Description of 45 translationese features and their extraction specifics for English, German and Russian used in LREC-2020 research

For results, refer to: Kunilovskaya, Maria & Ekaterina Lapshinova-Koltunski. Lexicogrammatic translationese across two targets and competence levels. LREC-2020, Marseille, May 11-15, 2020.

The extraction code is available from https://github.com/kunilovskaya/translationese45

Generally, the list includes

- seven frequencies of morphological forms (two degrees of comparison, past tense, passive voice form, two non-finite forms of verb (infinitive and all participles, deverbal nouns) and finite form;
- three syntactic functions in addition to UD relations: various PoS in attributive function, marker of subordinate clause, copula verbs
- eleven syntactic features that have to do with sentence type and structure (simple sentences, number of clauses per sentence, sentence length, interrogative and negative sentences, types of clauses (relative and pied-piped subtype, clausal complement), correlative constructions, modal predicates
- seven frequencies for morphological categories (five types of pronominal function words, adverbial quantifiers, all nouns)
- two graph-based features: mean hierarchical distance and mean dependency distance
- five list-based features for semantic types of discourse markers
- seven of 17 Universal Dependencies relations, that are shown to be useful translationese indicators for English>Russian translation pair in previous research (Kunilovskaya, Kutuzov 2018)[1]. These include such dependencies as adjectival clause, auxiliary, passive voice auxiliary, clausal complement, a word-introducer of a subordinate clause, subject of a passive transformation, ('acl', 'aux', 'aux:pass', 'ccomp', 'mark', 'nsubj:pass', 'parataxis', 'xcomp'). The values of these features are probability of occurrence of the relation in the sentence normalized to the number of sentences in the text
- two overall text measures of lexical density and variety

Note that for each language in this experiments we used the pre-trained model that returned most accurate results for our features and has the highest mean accuracy for UPOS, UFeats, Lemma and UAS reported at the respective UD page[2] among the available releases, at the time of writing it is 2.2 for English Web (English-EWT), 2.0 for German GSD, 2.3 for Russian-SynTagRus treebanks. The respective models performance rangers from 90% to 97% for UPOS and from 80%-94% for UFeats, with overall Unlabelled attachment score (UAS) of 74-89% for the three languages involved.

The extraction of discourse markers (including the four semantic groups of connectives and the markers of epistemic stance) is based on search lists for each of the project languages. The lists were initially produced independently from grammar reference books, dictionaries of function words and relevant research papers and then verified for comparability. Following Fraser (2006) discourse markers are treated functionally and include items of various morphological and structural types (conjunctions, adverbs, particles, parenthetical phrases). Each item on the lists was tested against the research data to take decisions in case of multifunctional items and to exclude homographs. Though most items on the lists are set phrases, we allowed for possible lexical and structural variability. We also used orthography and punctuation to disambiguate our DMs. Correlative conjunctions are left out for now.

1 Kunilovskaya, M., & Kutuzov, A. (2018). Universal Dependencies-based syntactic features in detecting human translation varieties. Proceedings Ofthe 16th International Workshop on Treebanks and Linguistic Theories (TLT16), 27–36.

2 http://ufal.mff.cuni.cz/udpipe/models#universal_dependencies_20_models

The classification follows the descriptions in Halliday and Hasan (1976)[3] and in Biber, et al (1999)[4]. Table 1 has the number of items in each group and the number of the items attested in our data as well as examples.

Table 1. Number of listed / attested discourse markers by category for each of the project languages and top most frequent items

	English	German	Russian
Additive and and or conjunctions are excluded to make their own feature	52 / 43	78 / 39	52 / 47
	Also, such as, for example, not only, for instance / in particular, moreover, in other words, namely	Auch, sowie, nicht nur, dafür, weiterhin / nämlich, das heißt, Dabei, übrigens, in Bezug auf, in gleicher Weise	Также, при этом, например, кроме того, в частности, к тому же, на самом деле, а именно, иными словами, точнее
Adversative but conjunction is excluded to make its own feature	46 / 35	88 /43	34 / 31
	Still, however, rather than, instead, though, on the other hand, in fact, despite	Jedoch, doch, allerdings, dennoch, gegenüber, außer in der Tat, eigentlich, trotz, umgekehrt, wogegen, einerseits, trotzdem	Однако, хотя, впрочем, правда, несмотря на, в отличие от, вместе с тем, всё-таки, но на самом деле, наоборот
causative	42 / 20	89 / 37	49 / 43
	Because, so, due to, so that, therefore, as a result, after all, for this reason, consequently	Damit, da, weil, denn, daher, deshalb, sodass, weshalb, auf diese Weise, als Folge, in diesem Fall, weswegen, wobei	Потому, поэтому, поскольку, ведь, так, в результате, ради того , чтобы, затем, что, получается, в этом случае, в связи с тем
temporal and sequential	110 / 76	147 / 75	48 / 41
	While, since, soon, and then, eventually, Further, Anyway, thus, at the same time, ultimately, meanwhile	Dann, also, nun, seit, dabei, bereits, gleichzeitig, zunächst, schließlich, später, zugleich, zweitens, damals, indes	Пока, наконец, затем, в целом, в то время, как, в заключение, в конце концов, во-первых, в то же время
epistemic markers + for EN introductory epistemic sentences with 1st person subjects in present tense ex. I/we think, I/we am/are (un)convinced/sure	64 / 45	74 / 40	86 / 78
	Really, at least, perhaps, of course, probably, in any case, for sure, in reality, no doubt, arguably, clearly, indeed	Natürlich, vielleicht, wahrscheinlich, sicher, möglich, wirklich, notwendig, möglicherweise, offensichtlich, überzeugt, meinen, in jedem Fall	Конечно, возможно, может быть, действительно, говорят, на мой взгляд, якобы, полагаю, по сути, в любом случае, кажется, бесспорно

3 Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English. Longman.

4 Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). Longman grammar of spoken and written English (Vol. 2). MIT Press. pp. 853-856

Normalization measures

We use two norms to make features comparable across different-size corpora, depending on the nature of the feature. Most of the features, including all types of discourse markers, negative particles, passives, types of verb forms, relative clauses, correlative constructions, adverbial clauses introduced by pronominal adverbs coordinating and subordinating conjunctions, simple sentences, number of clauses per sentence, are normalized to the number of sentences (30 features). Such features as personal, possessive and other noun substitutes, nouns, adverbial quantifiers, determiners are normalized to the running words (6 features). Counts for syntactic relations are represented as probabilities, normalized to the number of sentences (7 features). Some features have their own normalisation basis: comparative and superlative degrees are normalized to the total number of adjectives and adverbs, nouns in the functions of subject, object or indirect object are normalized to the total number of these roles in the text.

Description of all features (in the alphabetic order)

Acl

finite and non-finite clausal modifier of noun (adjectival clause), including relative clauses as a subtype (used only in EN and RU); extraction is based on UD default annotation

the person showing (acl) her around
help people do something to overcome (acl) it
as a language learned (acl) all over
it is a must that we move (acl:relcl) to English
an denen sich die Gliedmaßen der Kaulquappe herausbilden (acl)
die das Tier mit den Kiemen aufnimmt (acl)
die es erst noch zu erforschen (acl) gilt
мысль о том, что локомотивом роста должен (acl) быть
нет сомнений, что в этой стране не хотят (acl) повторять ошибок
людей, следящих (acl) за политикой
кино, где звезда плавает (acl:relcl) в трёх бассейнах

addit

additive connectives; cumulative frequency of the list items normalized to the number of sentences; see description in Table 1

advers

adversative (contrastive) connectives; cumulative frequency of the list items normalized to the number of sentences; see description in Table 1

attrib

adjectives and participles functioning as attributes; all words tagged as ADJ or VerbForm=Part with the amod dependency to their head

the rising sun
the coloured face

aux

auxiliary verbs; extraction is based on UD default annotation

aux:pass

auxiliary verbs in passive forms; extraction is based on UD default annotation

but

contrastive coordinating conjunction 'but' ('aber'/'но'), if not followed but 'also'/'auch'/'и', 'также' and not in the absolute sentence end

caus
causative connectives; cumulative frequency of the list items normalized to the number of sentences; see description in Table 1

ccomp
clausal complement as annotated in UD

help people to do (ccomp) smth
не ожидали, что придет

cconj
coordinating conjunctions: lemmas in 'and', 'or', 'both', 'yet', 'either', '&', 'nor', 'plus', 'neither', 'ether' ('und', 'oder', 'aber', 'sondern', 'sowie', 'als', 'wie', 'doch', 'sowohl', 'denn', 'desto', 'noch', 'weder', 'entweder', 'bzw', 'beziehungsweise', 'weshalb', 'und/oder', 'ob', 'woher', 'wenn', 'jedoch', 'wofür', 'insbesondere', 'obwohl', 'um' / 'и', 'а', 'но', 'или', 'ни', 'да', 'причем', 'либо', 'зато', 'иначе', 'только', 'ан', 'и/или', 'иль') tagged CCONJ. Lists are used to filter out noise.

comp
comparative degree of comparison for adjectives and adverbs; synthetic forms are extracted based on the tag Degree=Comp, while analytical forms are counted as adjectives and adverbs with a dependent 'more/mehr/более' and 'больший' for Russian

copula
copula verbs; lemmas of 'be', 'sein', 'быть', 'это' that have a 'cop' relation to their head, excluding constructions with there as head for English

correl
correlative constructions of all types, where a demonstrative PRON ('those', 'such', 'der', 'welch', 'was', 'wer', 'тот', 'такой', 'то') is syntactically or semantically connected to subsequent CONJ. In English they make a subset of relative clauses; in DE/RU they can also be a subtype of a clausal complement;

Trotzdem sorgen beide dafür , daß der Nucleus accumbens mit Dopamin überflutet wird , beziehungs- weise sie täuschen dort eine Dopaminschwemme vor (siehe Kasten S. 40) . Ich hörte weiter darüber , daß die Wirkung verbunden sei mit heftigem Erbrechen und Durchfall .

demdets
pronominal determiners; lemmas in the function ‘det’ from the lists 'this', 'some', 'these', 'that', 'any', 'all', 'every', 'another', 'each', 'those', 'either', 'such' ('dies', 'alle', 'jed', 'einige', 'solch', 'viel', 'ander ', 'jen', 'all', 'irgendwelch', 'dieselbe', 'jeglich', 'daßelbe', 'irgendein', 'diejenigen' / 'этот', 'весь', 'тот', 'такой', 'какой', 'каждый', 'любой', 'некоторый', 'какой-то', 'один', 'сей', 'это', 'всякий', 'некий', 'какой-либо', 'какой-нибудь', 'кое-какой')

deverbals
deverbal nouns, names of processes, actions, states. The extraction for EN/DE accounts for affixation (with most productive -ment, -tion/ -ung, -tion) and conversion as types of derivation. In the first case the output is filtered with an empirically-driven stoplist. Converted nouns are counted from a list of true procedural nouns that were not fully substantivised. To produce this list we looked through the nounal occurrences of lemmas that also appear as verbs and filtered out items that prevail in their fully substantivised lexico-semantic variants in our data (such as design, set, measure, mark, press, stick, cross, trap, handle). For Russian with extracted nouns in 'тие', 'ение', 'ание', 'ство', 'ция', 'ота' and employed a 150-items long stoplist to exclude fully

substantivised words such as собрание, месторождение, министерство, телевидение, творчество, решение.

epist
epistemic stance discourse markers; cumulative frequency of the list items normalized to the number of sentences; see description in Table 1

finites
verbs in finite form; for EN/RU we rely on UD morphological feature VerbForm=Fin
for DE we exclude False Positives after aux and modals and include False Negatives erroneously tagged VerbForm=Part (ex. Mit dieser Frage verlassen (Part!?) wir (plural nsubj) das Gebiet zellinterner Veränderungen und wenden uns)

indef
noun substitutes, i.e. pronouns par excellence, of indefinite, total and negative semantic subtypes, if token in 'anybody', 'anyone', 'anything', 'everybody', 'everyone', 'everything', 'nobody', 'none', 'nothing', 'somebody', 'someone', 'something', 'elsewhere', 'nowhere', 'everywhere', 'somewhere', 'anywhere'. For German: list items tagged PronType=Ind ('etwas', 'irgendetwas', 'irgendwelch', 'irgendwas', 'jedermann', 'jedermanns', 'jemand', 'alles', 'niemand', 'nichts', 'irgendwo', 'manch'). For Russian: 'когда', 'где', 'куда', 'откуда', 'отчего', 'почему', 'зачем' if tagged PRON; words with -то|-нибудь|-либо, except starting with 'какой'; and items from 'кто-кто', 'кого-кого', 'кому-кому', 'кем-кем', 'ком-ком', 'что-что', 'чего-чего', 'чему-чему', 'чем-чем', 'куда-куда', 'где-где'

infs
infinitives: all cases of a verb form tagged 'VerbForm=Inf' with a dependent to/zu particle and cases of true bare infinitive, excluding after modal verbs and 'have to', 'going to' and modal adjectival predicates, but including cases after 'help', 'make', 'bid', 'let', 'see', 'hear', 'watch', 'dare', 'feel', 'have' ('hören', 'sehen', 'spüren', 'lassen', 'gehen', 'bleiben', 'helfen', 'lehren'). For Russian all occurences of verb forms with the feature 'VerbForm=Inf' except after modal predicates and with the dependent 'быть' to exclude future forms ("отношения будут ухудшаться").

lexTTR
lexical type to token ratio: ratio of PoS disambiguated content words types (look_VERB vs look_NOUN) to their tokens. Content words include lemmas in 'ADJ', 'ADV', 'VERB', 'NOUN' part of speech categories.

mdd
mean dependency distance(MDD, aka comprehension difficulty) as “the distance between words and their parents, measured in terms of intervening words.” (Hudson 1995 : 16)[5]

mhd
mean hierarchical distance(MHD, aka production (speaker’s difficulty) as the average value of all path lengths travelling from the root to all nodes along the dependency edges (Jing, Liu 2015 : 164)[6]

mpred
modal predicates; for English all verbs tagged as 'MD' in XPOS, except 'will'/'shall', constructions with have-to-Inf and all adjectival modal predicates (predicative from a list of 17 items such as impossible, likely, sure with a dependent 'AUX'). For German: lemmas in 'dürfen', 'können',

5 Hudson, R. (1995). Measuring syntactic difficulty. Manuscript, University College, London.

6 Jing, Y., & Liu, H. (2015). Mean Hierarchical Distance Augmenting Mean Dependency Distance. Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), 161–170.

'mögen', 'müssen', 'sollen', 'wollen' or 'VM' in XPOS and all adjectival modal predicates formed with 15 listed adjectives (Selbstreparatur ist immerhin möglich., Es ist klar , dass ..). For Russian: lemma 'мочь', lemma 'следовать' with a dependent infinitive, three modal adverbs ('можно', 'нельзя', 'надо') and adjctives from the modal predicative list of 11 in the short form 'Variant=Short' (должен, способный, возможный)

mquantif
adverbial quantifiers; listed lemmas tagged ADV. For German we filter out tagging errors by excluding cases when the lemma depends on a noun. For Russian we additionally provide for functionally similar non-adverbial quantifiers such as 'еле', 'очень', 'вшестеро', 'невыразимо', 'излишне', 'еле-еле', 'чуть-чуть', 'едва-едва', 'только', 'капельку', 'чуточку', 'едва'. The support lists include EN: 37 items (barely, completely, intensely, almost), DE: 67 items (mässig, leidlich, ziemlich, völlig, ganz), RU: 80 items (абсолютно, полностью, сплошь, необыкновенно, достаточно, совершенно, невыносимо, примерно).

neg
negative particles or main sentence negation: counts of lemmas in 'no', 'not', 'neither' ('kein', 'nicht'/'нет', 'не')

nnargs
core verbal arguments represented by nouns or proper names; ratio of nouns and proper names in the functions of 'nsubj', 'obj', 'iobj' to the count of these functions

nsubj:pass
subjects of verbs in the passive voice; ; extraction is based on UD default 'nsubj:pass' annotation

numcls
number of clauses per sentence; number of relations from the list 'csubj', 'acl:relcl', 'advcl', 'acl', 'xcomp', 'parataxis' annotated in one sentence

passives
    bypassives
    passive constructions with expressed agentive role; all verbs tagged Voice=Pass with an 'obl' dependent in ['NOUN', 'PRON', 'PROPN'] having a dependent by (von) ADP. For Russian the 'obl' dependent is required to be have Case=Ins and Animacy=Anim in morphological features, filtering out 'образ', 'лето', 'осень', 'зима','весна', 'утро', 'вечер', 'ночь'
    shortpassives
    all passives without the explicit agent introduced by by or von prepositions and Russian equivalents (see bypassive). For English it is a verb with the annotated Voice=Pass feature and a dependent aux:pass. For German we also include ‘passive-like’: lassen sich + VerbForm=Inf. For Russian we account for both morphological forms (война велась, политика была направлена) and for semantic passive (стадион возводят на новом месте, во Владикавказе ему готовят радушную встречу)

parataxis
asyndatically connected coordinated clauses (often direct speech or clauses joined “:” or a “;” as well as parenthetical clauses); extraction is based on UD default annotation

pied
correlative constructions with displaced (pied-piped) preposition:
    technology for which Sony could take credit
    speech in which he made this argument

Man geht in der Biologie davon aus , daß die Entwicklung
die Zelle , aus der ein Ei hervorgeht
о таком, о каком вы не слыхали
скандал , в котором; трагедии , с которыми, в той конструкции , в какой она

possdet
possessive pronouns; for EN/DE lemma in 'my', 'your', 'his', 'her', 'its', 'our', 'their'/ 'mein', 'dein', 'sein', 'ihr', 'Ihr|ihr', 'unser', 'eurer' tagged 'DET', 'PRON' and 'Poss=Yes'. For Russian lemma in 'мой', 'твой', 'ваш', 'его', 'ее', 'её', 'наш', 'их', 'ихний', 'свой' tagged DET

ppron
personal pronouns; tokens tagged PRON, Person= that do not have Poss=Yes feature and are on the list 'i', 'you', 'he', 'she', 'it', 'we', 'they', 'me', 'him', 'her', 'us', 'them' ('ich', 'ihr', 'du', 'er', 'sie', 'es', 'wir', 'mich', 'mir', 'dich', 'dir', 'ihm', 'ihn', 'uns', 'ihnen' / 'я', 'ты', 'вы', 'он', 'она', 'оно', 'мы', 'они', 'меня', 'тебя', 'его', 'её', 'ее', 'нас', 'вас', 'их', 'неё', 'нее', 'него', 'них', 'мне', 'тебе', 'ей', 'ему', 'нам', 'вам', 'им', 'ней', 'нему', 'ним', 'меня', 'тебя', 'него', 'мной', 'мною', 'тобой', 'тобою', 'Вами', 'им', 'ей', 'ею', 'нами', 'вами', 'ими', 'ним', 'нем', 'нём', 'ней', 'нею')

pverbals
participles: for English all occurrences of 'VerbForm=Part' or 'VerbForm=Ger' not in attributive function 'amod' or part of an analytical form; for German all 'VerbForm=Part' not in the attributive function and 'ADJD' ending in 'd' in the 'advmod'/'acl' functions (Der Hund stand bellend (ADJ, 3, advmod) am Fenster.). For Russian 'VerbForm=Part' not in the short form and not in the attributive function, without a dependent auxilary, and 'VerbForm=Conv' without dependent auxiliary

relativ
all relative clauses, including correlative constructions and pied-piping construction. Extraction is based on affirmative sentences only: EN: words 'which', 'that', 'whose', 'whom', 'what', 'who' tagged as PRON', excluding cases when relative PRON has a dependent preposition and follows its head (ex. But we will return to that (PRON) later). DE: relative pronouns ('der', 'welch', 'was', 'wer') tagged as PRON and words with wo- of morphological category 'PronType=Int,Rel', if there is comma in the its left window of size 3 (-3:0 window). RU: 'который', 'что', 'кто', 'какой' and a comma in the left window of 3

sconj
subordinating conjunctions: lemmas in 'that', 'if', 'as', 'of', 'while', 'because', 'by', 'for', 'to', 'than', 'whether', 'in', 'about', 'before', 'after', 'on', 'with', 'from', 'like', 'although', 'though', 'since', 'once', 'so', 'at', 'without', 'until', 'into', 'despite', 'unless', 'whereas', 'over', 'upon', 'whilst', 'beyond', 'towards', 'toward', 'but', 'except', 'cause', 'together' ('daß', 'wenn', 'dass', 'weil', 'da', 'ob', 'wie', 'als', 'indem', 'während', 'obwohl', 'wobei', 'damit', 'bevor', 'nachdem', 'sodass', 'denn', 'falls', 'bis', 'sobald', 'solange', 'weshalb', 'ditzen', 'sofern', 'warum', 'obgleich', 'zumal', 'sodaß', 'aber', 'wenngleich', 'wennen', 'wodurch', 'wohingegen', 'ehe', 'worauf', 'seit', 'inwiefern', 'anstatt', 'der', 'vordem', 'insofern', 'nahezu', 'wohl', 'manchmal', 'weilen', 'weiterhin', 'doch', 'mit', 'gleichfalls' / 'что', 'как', 'если', 'чтобы', 'то', 'когда', 'чем', 'хотя', 'поскольку', 'пока', 'тем', 'ведь', 'нежели', 'ибо', 'пусть', 'будто', 'словно', 'дабы','раз', 'насколько', 'тот', 'коли', 'коль', 'хоть', 'разве', 'сколь','ежели', 'покуда', 'постольку') tagged SCONJ. Lists are used to filter out noise.

simple
simple sentence; a sentence where no words have relations from the list 'csubj', 'acl:relcl', 'advcl', 'acl', 'xcomp', 'parataxis'

sup
superlative degree of comparison for adjective and adverbs; synthetic forms are extracted based on the tag Degree=Sup, while analytical forms are counted as adjectives and adverbs with a dependent most/наиболее/самый and for Russian words starting with 'наи-' with the exception of a few homonymous adverbs ('наискосок')

tempseq
temporal and sequential connectives; cumulative frequency of the list items normalized to the number of sentences; see description in Table 1

whconj
adverbial clause introduced by a pronominal ADV on the lists 'when', 'where', 'why' ('wann', 'wo', 'warum' / 'когда', 'где', 'куда', 'откуда', 'отчего', 'почему', 'зачем')

xcomp
a predicative or clausal complement without its own subject, annotated after phasal verbs (started to sing), in case of infinitive constructions (asked me to leave), etc.; extraction is based on UD default annotation