Measuring Translationese across Levels of Expertise: Are Professionals more Surprising than Students?

Yuri Bizzoni* **Ekaterina Lapshinova-Koltunski***
Saarland University, Campus A2.2, Saarbrücken, Germany
yuri.bizzoni@uni-saarland.de e.lapshinova@mx.uni-saarland.de

Abstract

The present paper deals with a computational analysis of translationese in professional and student English-to-German translations belonging to different registers. Building upon an information-theoretical approach, we test translation conformity to source and target language in terms of a neural language model's perplexity over Part of Speech (PoS) sequences. Our primary focus is on register diversification vs. convergence, reflected in the use of constructions with a higher vs. lower perplexity score. Our results show that, against our expectations, professional translations elicit higher perplexity scores from the target language model than students' translations. An analysis of the distribution of PoS patterns across registers shows that this apparent paradox is the effect of higher stylistic diversification and register sensitivity in professional translations. Our results contribute to the understanding of human translationese and shed light on the variation in texts generated by different translators, which is valuable for translation studies, multilingual language processing, and machine translation.

1 Introduction

Translationese is a set of linguistic patterns that tell translations apart from texts originally written in the same language and that make translations stylistically more similar to each other than original texts tend to be. While translationese was extensively discussed in the area of corpus-based translation studies and machine translation (MT) (Zhang and Toral, 2019; Graham et al., 2020, among others), there are relatively few computational studies that focus on the varying amount of translationese characterizing different kinds of written translation (see Section 2.2 below). This study focuses on the relation between translators' level of expertise and translationese throughout different registers. If we can connect translationese at least partly to the translator's experience, we can expect to find different degrees of translationese between student and professional translations. As translationese is probabilistic in nature (Toury, 2004), we use a framework that enables a probabilistic design of language use in the form of a language model. We model language conventions in terms of grammatical structures represented by PoS sequences through Long Short-Term Memory (LSTM), a recurrent neural network architecture, using monolingual corpora of non-translations in both source and target language as a training set. We then test how students' and professionals' translations conform to linguistic conventions using our models' perplexity scores. Through this approach, we aim at testing two related hypotheses:

Hypothesis 1 Overall, we can expect professional translators to be more efficient at reproducing the patterns of their target language. If this is the case, we would expect professional translations to elicit lower perplexity scores from the target language model than from the model of the source language.

Hypothesis 2 On the other hand, it is possible that students converge more on standard patterns: due to their lack of expertise, they might have lower register sensitivity, and thus they could be less bold and more repetitive in their use of grammatical constructions. A higher value of perplexity for a register means a less usual (hence, more perplexing) order of PoS with respect to a reference corpus,

*Both authors contributed equally.

and hence a more distinct register. So a higher level of convergence would result in more homogeneous surprisal values across registers.

We then compare the results of our perplexity measures with the distribution of different PoS patterns across registers to qualitatively analyze translation divergence in the data.

We organized the remainder of the paper as follows: Section 2 introduces the main concepts we are developing our research on and provides an overview of the related work, Section 3 includes details on the data and methods used in the analyses. We show the results of the analyses in Section 4 and 5, interpret them, and conclude in Section 6.

2 Main Concepts and Related Work

2.1 Distinctive features of translations

As we mentioned above, translationese is related to a set of distinctive linguistic features that make translations differ from non-translations (Gellerstam, 1986; Baker, 1996; Toury, 1995). Translationese appears to be a ubiquitous phenomenon, and happens in different forms both in human and machine translations (Graham et al., 2020; Zhang and Toral, 2019; Bizzoni et al., 2020). Automatic classifications of texts into translations and non-translations usually operationalize translationese as a combination of lexico-grammatical and textual features of different kinds (Baroni and Bernardini, 2006; Laippala et al., 2015; Volansky et al., 2015; Rabinovich et al., 2017). The number of such features, as well as their designation, varies across translation studies. The following macro-categories are relevant for our study: **shining through** – translations reproducing patterns typical of the source language instead of following the target language conventions (Teich, 2003)[1]; **normalization** – translations conforming to patterns and practices which are typical of the target language (Baker, 1996), and **convergence** – the tendency of translated language to be more homogeneous in terms of the distribution of language patterns (Laviosa, 2002). In our definition of convergence, we follow the study by Kruger and van Rooy (2012) who analyze it across registers and conceptualize it as a form of register insensitivity. The main idea is that translations show less variation, which reduces the distinctness of various registers. While in this sense, there might be no "perfect" translation (no translation completely indistinguishable from comparable originals), we are interested in the degree to which professional and non-professional translators are sensitive to register.

We can observe translationese at the lexical level, i.e., displaying less lexical and semantic diversity than the original (Baroni and Bernardini, 2006; Bizzoni and Teich, 2019), and at the grammatical level, i.e., using more typical syntactic construction instead of unusual ones (Ilisei et al., 2010; Volansky et al., 2015). In our analyses, we test translation conformity with either the target or the source language through language models' perplexity scores measured over PoS sequences which represent grammatical level.

2.2 Translationese and translation expertise

Computational analyses of professional and novice translations are based on the assumption that translations of different levels of expertise manifest translationese to different degrees. Redelinghuys (2016) compare non-translated English texts with translations by experienced and inexperienced translators in terms of frequencies of features like conjunctive markers, standardized type-token ratio, and word length, performing a univariate analysis for individual features. Kunilovskaya and Lapshinova-Koltunski (2020) report on two separated translationese effects and find a correlation between the levels of expertise and types of the detected effects. However, they ignore register differences. Lapshinova-Koltunski (2020) shows in her analyses of the same translation dataset we are using in our work that there are register-specific effects on the normalization and shining through for professional and student translations. Her results are based on such measures as distribution of content and grammatical words, nominal and verbal categories, various types of pronouns a.o., however, and do not provide any significant differences between student and professional translators. Corpas Pastor et al. (2008) and Ilisei (2012) use supervised machine learning techniques to distinguish between non-translations in Spanish and English-Spanish translations by professionals and students, investigating the validity of the translation universal of convergence. However, their definition of convergence is different from ours – they define this as the similarity between texts trans-

[1] Shining through is related to the law of interference, according to which phenomena of the make-up of the source text tend to get transferred to the target text (Toury, 1995).

lated by translators of different proficiency levels and do not find significant differences between student and professional translations in terms of the features applied. We relate our analysis to the study by Martínez and Teich (2017) who also apply a probabilistic approach to study differences in the lexical choices by professional and student translators related to source-dependent and target-dependent translationese. Rubino et al. (2016) also use surprisal measures based on lexical, PoS, and syntactic patterns to analyze translationese in a dataset containing human translations with different levels of expertise, focusing on the automatic separation of non-translations from translations. This work addresses convergence as the proximity of two translation variants (professional and student). Register awareness is one of the critical elements of translation expertise (Olohan, 2015) — for example, the mentioned study by Redelinghuys (2016) show inexperienced translators to be more repetitive when translating creative writing than popular texts, which points to their practice in the academic context of translator training. A recent study by Popović (2020) explores differences between texts translated by professional translators, crowd contributors, and translation students, showing their impact on machine translation evaluation. This study suggests that it is crucial for machine translation evaluation to understand the factors influencing human translation variation, especially when we compare human and machine translation quality.

2.3 Register in translation

Our definition of register relies on variational linguistics (Biber, 1995; Halliday, 1985). Variation across registers is linked to the distribution of linguistic patterns in different contexts: register diversification represents distinctive distributions of linguistic patterns, as compared to the use of these patterns in other contexts (Biber et al., 1998, 13). Register variation has also been an object of analysis in translations. Kruger and van Rooy (2012) state that translationese is subject to the influence of register and Neumann (2013) demonstrates the degree to which translations get adapted to the requirements of different registers in English and German. Her feature set inspired the study by Evert and Neumann (2017) who detect similarities between register and translationese features. Lapshinova-Koltunski (2017) analyses the interaction between register and translation method (human vs. machine), also paying attention to the differences between professional and novice translators. Lapshinova-Koltunski and Zampieri (2018) automatically discriminate registers and translation methods using part of speech n-grams. They show that it is harder to automatically differentiate between translation methods than between registers. This means that register diversification prevails over translation method diversification. This also points to a convergence between translations, which is of interest in our work. However, this convergence is related to the translations produced with different methods and not to the reduced distinction of various registers in favor of a more neutral "middle" register, as defined by Kruger and van Rooy (2012) and pursued in our work.

3 Research Design

3.1 Data

We use a dataset of English and German texts exported from two corpora. We derived English originals (EO), their translations into German by professionals (PT), as well as comparable German non-translations (GO) from the CroCo corpus (Hansen-Schirra et al., 2012). The non-professional translations (ST) for the same English sources as in CroCo were produced by students of translation and come from the corpus VARTRA (Lapshinova-Koltunski, 2013)[2]. In this way, both professionals' and students' translations have the same sources and represent translation variants of the same original texts. Our dataset covers seven registers: political essays (ESSAY), fiction (FICTION), manuals (INSTR), popular-scientific articles (POPSCI), letters to shareholders (SHARE), prepared political speeches (SPEECH), and tourism leaflets (TOU). The English sources and the comparable German non-translated texts used for training our language models cover the same registers. In Table 1, we provide details on the size of the data under analysis.

To ensure the comparability of the models' results in the source and the target languages, we use the Universal PoS tagset (Petrov et al., 2012). All texts in the data were automatically tokenized, lemmatized, and annotated with part of speech infor-

[2]We define professional translators as experts who have a good degree of experience in translating, mostly specializing in their areas, whereas students are trainees who have no or little experience in translation. While the two groups inhabit a continuum, we are happy with a binary division

	EO	GO	ST	PT
ESSAY	35 238	36 162	16 295	35 865
FICTION	37 019	36 913	12 755	37 953
INSTR	35 668	36 562	20 816	35 342
POPSCI	35 668	36 321	23 369	33 880
SHARE	36 437	35 517	25 630	36 810
SPEECH	35 223	35 769	24 999	36 377
TOU	35 981	36 564	20 358	34 139
TOTAL	251 894	253 862	144 222	250 366

Table 1: Dataset size in tokens.

mation based on the Universal Dependency framework (Straka and Straková, 2017). The accuracy of automatic annotation of the respective models for universal parts of speech is 90.5% for German and 94.5% for English[3]. Naturally, our PoS taggers can make mistakes, and it is conceivable that this margin of error might bring them to label some unusual sequences of words with more conventional, albeit erroneous, tags. Even if this anomaly were to happen, we find that it could not account for the differences we observe among our corpora since it would affect all texts similarly, and it would at worst slightly reduce their differences rather than magnify them.

While the amount of data is small for neural network training, it is essential to remember that since we are using a universal tagset, its vocabulary size is tiny: 15 parts of speech in total. This vocabulary size keeps the complexity of the learning process drastically lower than that of word sequence modeling and it allows our network to model small data well enough to display systematically lower perplexities when presented with unseen documents from the corpus on which it was trained (see for example Table 2).

3.2 Perplexity

We train two standard one-layered LSTM language models (LM) of 50 cells [4] on the PoS sequences of 80% of the whole English and German non-translations respectively and measure their perplexity on professional translations, student translations, and originals. Even with small training data, our language models display lower perplexities for unseen instances of the originals from which we sampled the training set (see Table 2 in Section 4 below) than for translations in the same language.

We try two training sets: in the first case, we train LMs on the unweighted, randomly sampled 80% of the corpus. In the second case, we train our language models on a representative sample that respects the whole corpus' genre percentages. In this way, we try to prevent domain bias from distorting our results in the test phase. In both cases, we test our models on unseen PoS sequences from originals or translations and analyze their average perplexity – a measure of how well a probability distribution predicts a sample as defined in (1), where $\{w_1, \ldots, w_T\}$ is held out test data that provides the empirical distribution $q(\cdot)$ in the cross-entropy formula given in (2) and $p(\cdot)$ is the language model (LM) estimated on a training set.

$$PP = 2^{\tilde{H}_r} \quad \text{where} \quad \tilde{H}_r = -\frac{1}{T}\log_2 p(w_1, \ldots, w_T) \quad (1)$$

$$\tilde{H} = -\sum_x q(x) \log p(x) \quad (2)$$

In this way, perplexity delivers a measure similar to surprisal in Information Theory (Shannon, 1948), according to which language items with high surprisal/ low predictability convey more information than items with low surprisal/ high predictability in context. Our analyses use neural language models' average perplexity for the PoS n-grams in all the subcorpora under analysis. In terms of n-gram language models, predictability in context means p(unit|context), where context is the preceding context of n-1 words. A higher value of perplexity for a text means high surprisal/low predictability and, hence, an order of PoS sequences unusual for a reference corpus. We run perplexity-based tests for the remaining 20% of the non-translations and on both student and professional translations. We expect that the *relative* perplexity of English-trained and German-trained models (independently from their baselines) can tell us something about grammatical translationese.

We expect low perplexity values on the monolingual data (e.g. German non-translations on German non-translations) and high perplexity values on cross-lingual data (e.g. German non-translations on the English model). We also expect translations to fall between the source and the target language. In this way, perplexity values for translated data should be higher than those of non-translations within one language but lower than

[3]See `http://ufal.mff.cuni.cz/udpipe/models#universal_dependencies_20_models` for details.

[4]We used Keras 2.0.9 (Chollet et al., 2015) running on Tensorflow 1.10.0 (Abadi et al., 2016)

the cross-lingual values. We also expect perplexity values for the professional translations to be lower than for the student translations within one language, but higher when tested cross-lingually (Hypothesis 1).

In terms of register diversification in the translated data, the essential idea is that an LM trained on a diverse set of registers[5] will find, on average, a converging translation less perplexing, since it contains grammatical structures typical of what we could call "general language". Thus we expect higher perplexity values for registers characterized by a distinctive or creative use of language – i.e., fiction – and lower values for more conventionalized registers – such as instruction manuals. Convergence will result in the homogeneity of perplexity values across different registers. Here, we expect a higher homogeneity, and hence convergence, for students than professionals (Hypothesis 2).

3.3 Pattern analysis

In the last step, we compare our perplexity results with the distributions of PoS n-grams across registers and corpora. Distributions of different PoS n-grams should show whether professional and student translators tend to be more repetitive or more diverse in using typical structures while translating various target language registers. So, we run a comprehensive examination of how many distinct PoS patterns translators use in a given text portion. Since our data contains the same source texts (and thus the same source patterns), we can expect that the more perplexing specific translations are, the more diverse patterns they should be.

We analyze PoS pattern diversity – the number of different PoS n-grams used in each register by students and professionals, which shows how many different PoS patterns translators use in a given portion of text and determine whether professionals are more diverse in translating registers than students. If students have an accentuated tendency to converge, they should show less diversity than professionals, which is especially revealing given that both professionals and students are translating the same source text, starting from the same source-structures. For all analyses, we have studied the differences between our subcorpora with growing n-grams, moving from bigrams up to heptagrams. We ran them on the same amount of text for all subsets, thus down-sampling the professional and the original corpora.

4 Perplexity Score Analyses

4.1 Hypothesis 1

We report the perplexity scores not controlling for register in Table 2, which illustrates the results of the model performance on all the four subcorpora under analysis, as well as the results of the t-test showing that the models' differences in perplexity are all statistically significant.

	EO-LM	GO-LM	t-value	p-value
EO	8.88	15.08	-11.6	<0.001
GO	11.12	5.93	23.5	<0.001
ST	12.51	11.12	3.2	0.001
PT	11.36	14.39	-10.1	<0.001

Table 2: Perplexity of the English-trained (EO-LM) and the German-trained models (GO-LM) on EO, GO, ST, and PT along with the results of t-test (t and p-value).

As stated in Section 3.2, we expect lower perplexity values for the tests within the monolingual data samples than for the cross-lingual data samples. Our English model is more surprised seeing other English PoS n-grams than German seeing other German PoS n-grams (8.88 vs. 5.93), which might derive from a more significant variation of morpho-syntactic patterns in the English data. A sanity check on the n-gram distribution shows that in our data, English has more diversity than German in terms of language patterns: for the vast majority of n-grams selections, English appears to have a higher number of different structures than German, which could be justified by the analytical character of English if compared to German: English uses more prepositions and auxiliaries to build up various constructions, whereas German expresses the same meaning through morphological strategies (endings, suffixes) that are not captured by the PoS n-grams. It is interesting to see that English and German are not equally surprised by each other: the English model is less surprised to see German n-grams (11.12) than is the German model when seeing English n-grams (15.08) [6]. Taking the language status of English, this might be, on the one hand, surprising as English has much influence on the German language, which takes over

[5] we trained LMs on the texts of the target language corpus that represent all registers.

[6] The differences between the these distributions are statistically significant.

English structures (structural anglicisms). On the other hand, we can explain this difference again by the diversity of language patterns in the English data: we can expect a system modelled on English to be more used to structural change and, as such, less surprised by the new structures it encounters in German.

The English model is less perplexed by professional translations (11.36) than by non-professional ones (12.51). In this way, professionals seem to be closer to their source texts (interference). Student translations elicit a higher perplexity score (12.51), which indicates that they are even more surprising for the English model than the comparable German non-translations and translations by professionals, which indicates over-normalization – exaggerating the target language patterns as defined in Section 2.1. The German model's results reveal an opposite tendency: professional translations seem to be more perplexing to the German model than the student ones. The German model seems to be highly surprised by the PoS sequences in the professional translations. Interpreting this result in terms of translationese, such a high level of perplexity, not far from the perplexity reached by English data, could indicate a degree of interference in professionals. This tendency is against our expectations formulated in Hypothesis 1 in Section 1.

4.2 Hypothesis 2

In the next step, we look into perplexity scores controlling for register in order to analyze convergence. We summarize our results in Table 3. We used a mixed training set that included a balanced number of sentences from each domain to maximize the data's representativity.

	ST	PT	t-test	p-value
FICTION	11.41	12.74	**-5.6**	<.001
ESSAY	10.54	13.73	**-14.2**	<.001
POPSCI	10.20	10.50	**-1.6**	<.001
INSTR	8.59	9.63	**-5.2**	<.001
SHARE	12.65	13.23	-0.5	0.5
SPEECH	10.08	9.83	1.2	0.2
TOU	10.22	12.34	**-9.04**	.001
ALL	11.12	14.39	**-2.45**	0.01

Table 3: Perplexity of the German-trained model on ST and PT. We also report t-test and p value for each pair of distributions. We bolded the statistics that reject H_0 at the 0.05 significance level.

As seen from the table, all registers translated by professionals elicit higher scores than those translated by students, except for political speeches. However, the scores for this register, as well as those for letters to shareholders do not show a statistically significant difference between the two groups of translators. We interpret the lower scores of student translations as a reduced register distinction in favor of a more general language, which confirms our hypothesis that students are more repetitive in the language constructions they use. One of the reasons for this tendency could be that students tend to employ specific transfer patterns when translating from English into German, resulting in the frequent use of conventional structures and, consequently, a higher convergence of their translations. Another explanation could be that novice translators do not have enough knowledge about specific registers and various aspects of technical communication. Therefore, they translate different registers similarly, making them closer to the general language in German. Because they tend to repeat the same patterns for different registers, students seem less perplexing than professionals. We verify these assumptions in the experiments on pattern diversity in Section 5. We also compare the perplexity values across registers for both translation varieties. Using the scores in Table 3, we rank the registers for student and professional translations in Table 4.

We observe a very similar ranking for all registers in both translation variants. The only exception seems ESSAY, which is more distinct in professional translations and more conventionalized in student translations; one reason for this might be, as we will detail later, the over-normalization of other domains (i.e., FICTION) in student translations. The most conventionalized register in both translation varieties is INSTR. It is also interesting to see that the scores for registers translated by students are less variable than those for registers translated by professionals, which indicates register-related convergence in student translation.

5 Analysis of Pattern Diversity

Figure 1 illustrates the number of unique PoS n-grams used in the different registers of our German corpora by professionals (left graph) or students (right graph) – on the x-axis – as compared to the number of unique PoS n-grams used in the same registers by comparable German originals – on the y-axis.

PT	INSTR ⇒ SPEECH ⇒ POPSCI ⇒ TOU ⇒ FICTION ⇒ SHARE ⇒ *ESSAY*
ST	INSTR ⇒ SPEECH ⇒ POPSCI ⇒ TOU ⇒ *ESSAY* ⇒ FICTION ⇒ SHARE

Table 4: Register ranking according to their perplexity scores.

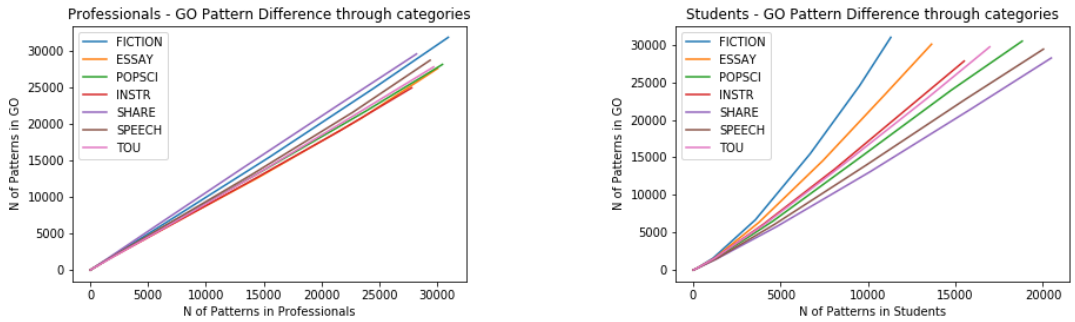

Figure 1: Differences between PoS n-grams in GO and PT (left side) and in GO and ST (right side), going from bigrams to heptagrams. For example, FICTION in GO has more than 30.000 different heptagrams, while FICTION in Students has about 10.000; instead, SPEECH progresses similarly for both categories through all ngrams, drawing a straighter line

We see from these graphs that professionals tend to have register-specific variations that are substantially similar to those of the equivalent originals, while students appear to be less diverse than both comparable originals and professionals, especially in more "creative" registers such as FICTION or ESSAY. Interestingly, the differences between professional and student translators (Figure 2) appear to be similar to those observed for ST and GO.

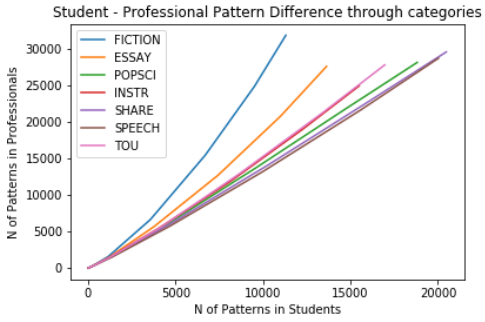

Figure 2: Differences between PoS patterns in the PT and the ST registers, going from bigrams to heptagrams, with heptagrams marking the end of each line.

It seems, overall, that the reason for the lower perplexity scores of the PoS-based language models for student translations is that students overnormalize their outputs, reusing fewer but more predictable structures. Professionals are more creative in their sentence structures: they are thus more perplexing for a general German model, but their behavior is, paradoxically, more similar to that of original writers. We illustrate the differences in language patterns discovered between student and professional translators with examples (1) and (2). For this, we pick exemplars for which students turn to be more repetitive than professionals while translating the same text. We illustrate the pattern NOUN-ADP-DET-NOUN-DET-NOUN-ADP in student translation in (1), whereas the ST version in (2) displays an example of the VERB-ADP-DET-NOUN-ADJ-PUNCT-SCONJ structure.

(1) a. *Seine Initialen, SR <...> waren in den Torbögen eingraviert und zogen sich durch das [Gebäude wie die Graffitimalereien der Gangs in] den Straßen der Stadt.*

b. *Und hier und da seine Initialen, SR <...> in Torbögen eingeritzt, [wie die Bandengraffiti] draussen auf der Strasse*

c. *And his initials here and there, SR <...> carved in archways [like the gang graffiti in] the streets outside.*

In (1-a), the student translator uses a complex nominal structure and adds some information not available in the source. The translation by a professional in (1-b) contains the same information as in the source (1-c) and a more lexically dense structure (*Bandgraffiti* vs. *Graffitimalereien der Gangs*).

(2) a. *Der Schweiß [lief an unserem Körper*

herunter, sodass] unsere T-Shirts an unsere Rücken klebten.

b. *Der Schweiß [lief uns am Körper runter, daß] uns die Hemden am Rücken klebten.*

c. *The sweat [came down our bodies and] plastered our shirts to our backs.*

Both translation varieties in example (2) convey the same information from the source sentence. However, the translation by a professional in (2-b) sounds more natural in German, whereas the student translation in (2-a) is closer to the source sentence. The direct object in the English in (2-c) cannot be directly transferred into German because of the restriction on semantic diversity of subjects and objects in German. The professional translator changes the direct object into a Dative+prepositional object, whereas the student uses just a prepositional object.

6 Conclusion and Outlook

In this study, we analyzed translationese in professional and student translations using a perplexity-based approach. We modelled the source and target grammatical patterns with an LSTM architecture and tested the conformity to the source and target language conventions of the two translation varieties through PoS perplexity. Despite the relative scarcity of our data, the small vocabulary of universal PoS allowed our LSTMs to learn the short and long-distance patterns well enough to display significantly higher perplexities when confronted with translations instead of original texts. Through this method, we found that, surprisingly, professional translators elicit higher perplexity scores from the target language model than students, which is against our first hypothesis. Nonetheless, in the analysis of convergence, we tested the extent to which professional and student translations of various registers conform to the target language model. We found more convergence in student than in professional translations, confirming our second hypothesis. We then tried to understand such results by analyzing PoS n-gram patterns in both translation varieties and conducting a qualitative analysis of translation divergence in the data. Overall, we found that such higher perplexities are an artifact of higher register variation in professional translations. We are not observing interference, but rather professionals' essential ability to be more daring with their language use. Student translators converge more, displaying a lower register sensitivity and a tendency to overuse the most general structures of the target language, while professional translators display more diversity and creativity in their structures, behaving in this way more similar to native writers.

The qualitative analysis of the examples suggest that the source of this diversity may originate from the cross-lingual differences between the source and the target languages: faced with a construction that has no direct or obvious equivalent in the target language, students might tend to choose less brilliant, more standard constructions across registers, whereas professionals might attempt to recreate the original domain's diversity. At the same time, we realize that our analyses may have some limitations. For instance, due to the absence of metadata in professional translations, we fail to control for individual variation in the data. For students, we know that the texts of various registers were sometimes translated by the same translators.

The results of our analyses provide an empirical contribution to the understanding of human translation. They show evidence of variation between texts generated by different translators in terms of language patterns and shed more light on the phenomenon of translationese. Studying variation in human translations of the same source texts across various registers is valuable for translation studies and multilingual language processing, especially for MT. As shown in Popović (2020), human translation variation plays a great role in MT evaluation.

In the future, we plan to deepen our understanding of how students over-normalize by aligning source and target texts, allowing for qualitative analyses of their translating behavior. We also want to explore whether other factors beyond the level of expertise influence translation convergence. Moreover, we would like to connect these results with the growing field of automatic translation quality estimation. Finally, although it is hard to find appropriate datasets containing comparable texts in terms of registers and different degrees of expertise, it would be interesting to expand this work on the opposite translation direction (German-English) and other language pairs to see if such tendencies are universally valid. Multilinguality would introduce more variance, and thus more factors to consider to avoid the risk of overclaiming and misunderstanding the complex phenomenon of translationese.

Acknowledgments

This work has been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 232722074–SFB 1102.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. TensorFlow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283.

Mona Baker. 1996. Corpus-based translation studies: The challenges that lie ahead. In H.L. Somers, editor, *LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, page 175–186. John Benjamins Publishing Company, Amsterdam.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

Douglas Biber, Susan Conrad, and Randi Reppen. 1998. *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge University Press, Cambridge.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? Comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290.

Yuri Bizzoni and Elke Teich. 2019. Analyzing variation in translation through neural semantic spaces. In *Proceedings of the 12th Workshop on Building and Using Comparable Corpora (BUCC) at RANLP-2019*.

François Chollet et al. 2015. Keras. `https://keras.io`.

Gloria Corpas Pastor, Ruslan Mitkov, Naveed Afzal, and Lisette Garcia-Moya. 2008. Translation universals: do they exist? a corpus-based and nlp approach to convergence. In *Proceedings of the LREC-2008 Workshop on Building and Using Comparable Corpora*, pages 1–7.

Stefan Evert and Stella Neumann. 2017. The impact of translation direction on characteristics of translated texts: A multivariate analysis for English and German. *Empirical Translation Studies: New Methodological and Theoretical Traditions*, 300:47.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In L. Wollin and H. Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Michael Alexander Kirkwood Halliday. 1985. *Spoken and Written Language*. Deakin University, Victoria.

Silvia Hansen-Schirra, Stella Neumann, and Erich Steiner. 2012. *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. de Gruyter, Berlin, New York.

Iustina Ilisei. 2012. *A machine learning approach to the identification of translational language: an inquiry into translationese*. Doctoral thesis, University of Wolverhampton.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. Identification of translationese: a supervised learning approach. In *Proceedings of CICLing-2010*, volume 6008 of *LNCS*, pages 503–511, Springer, Heidelberg.

Haidee Kruger and Bertus van Rooy. 2012. Register and the Features of Translated Language. *Across Languages and Cultures*, 13(1):33–65.

Maria Kunilovskaya and Ekaterina Lapshinova-Koltunski. 2020. Lexicogrammatic translationese across two targets and competence levels. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4102–4112, Marseille, France. European Language Resources Association.

Veronika Laippala, Jenna Kanerva, Anna Missil, Anna Missilä, Sampo Pyysalo, Tapio Salakoski, and Filip Ginter. 2015. Towards the Classification of the Finnish Internet Parsebank : Detecting Translations and Informality. In *Nodalida*. Linköping University Electronic Press, Sweden.

Ekaterina Lapshinova-Koltunski. 2013. VARTRA: A Comparable Corpus for Analysis of Translation Variation. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 77–86, Sofia, Bulgaria. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski. 2017. Exploratory analysis of dimensions influencing variation in translation: The case of text register and translation method. In Gert De Sutter, Marie-Aude Lefer,

and Isabelle Delaere, editors, *Empirical Translation Studies: New Methodological and Theoretical Traditions*, volume 300 of *TILSM series*, pages 207–234. Mouton de Gruyter. TILSM series.

Ekaterina Lapshinova-Koltunski. 2020. Tracing normalisation and shining through in novice and professional translations with data mining techniques. In Sylviane Granger and Marie-Aude Lefer, editors, *Translating and Comparing Languages: Corpus-based Insights*, volume 6 of *Corpora and Language in Use Proceedings*, pages 33–47. Presses universitaires de Louvain, Louvain-la-Neuve.

Ekaterina Lapshinova-Koltunski and Marcos Zampieri. 2018. Linguistic features of genre and method variation in translation: A computational perspective. In Th. Charnois, M. Larjavaara, and D. Legallois, editors, *The Grammar of Genres and Styles: From Discrete to Non-Discrete Units*, volume 320 of *TILSM series*, pages 92–117. Mouton de Gruyter.

Sara Laviosa. 2002. *Corpus-based Translation Studies, Theory, Findings, Application*. Rodopi, Amsterdam.

José Manuel Martínez Martínez and Elke Teich. 2017. Modeling routine in translation with entropy and surprisal: A comparison of learner and professional translations. In Larissa Cercel, Marco Agnetta, and Maria Teresa Amido Lozano, editors, *Kreativität und Hermeneutik in der Translation*. Narr Francke Attempto Verlag.

Stella Neumann. 2013. *Contrastive register variation. A quantitative approach to the comparison of English and German.* Mouton de Gruyter, Berlin, Boston.

Maeve Olohan. 2015. *Scientific and technical translation*. Routledge.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096.

Maja Popović. 2020. On the differences between human translations. In *Proceedings of EAMT-2020*, Lisboa, Portugal.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in Translation: Reconstructing Phylogenetic Language Trees from Translations. *Acl-2017*, pages 530–540.

Karien Redelinghuys. 2016. Levelling-out and register variation in the translations of experienced and inexperienced translators: a corpus-based study. *Stellenbosch Papers in Linguistics*, 45(0):189–220.

Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef van Genabith. 2016. Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of NAACL HT 2006*, pages 960–970, San Diego, California.

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Milan Straka and Jana Straková. 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.

Elke Teich. 2003. *Cross-Linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Berlin.

Gideon Toury. 1995. *Descriptive Translation Studies - and Beyond*, benjamins edition. John Benjamins Publishing Company.

Gideon Toury. 2004. Probabilistic explanations in translation studies: Welcome as they are, would they qualify as universals? In A. Mauranen and P. Kujamäki, editors, *Translation Universals: Do They Exist?*, Benjamins translation library, pages 15–32. J. Benjamins Publishing Company.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

Appendix A. Additional Figures

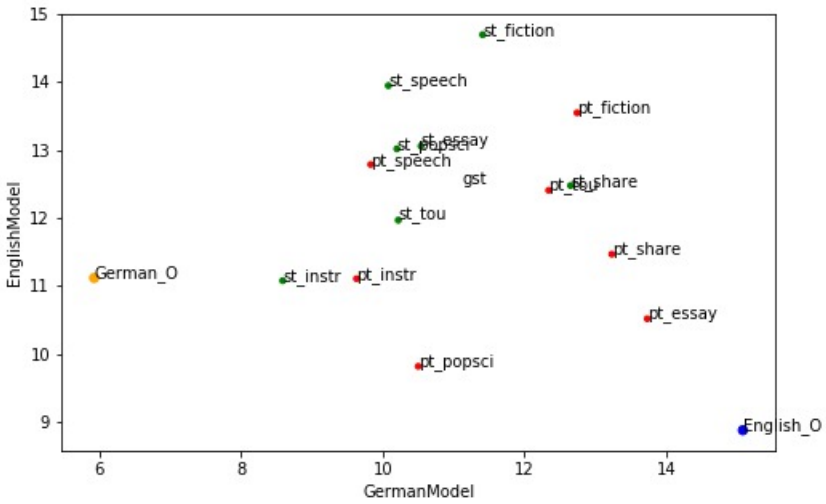

Figure 3: Perplexity of various registers of PT (red) and ST (green) for the English and German models, as well as general EO (blue) and general GO (yellow).

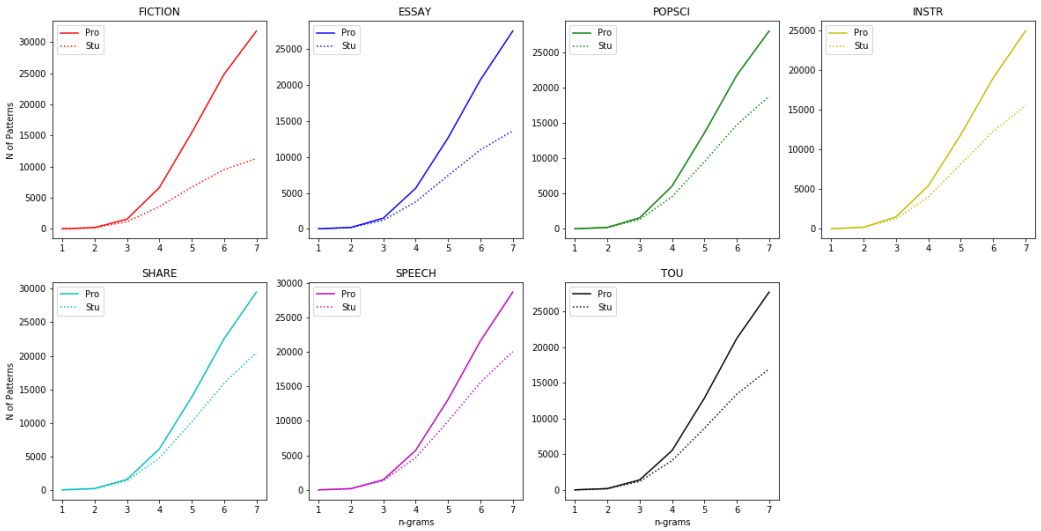

Figure 4: Number of different patterns used by students and professionals per each category, with growing n-gram length.