Using Machine Learning to Validate a Novel Taxonomy of Phenomenal Translation States

Michael Carl
Kent State University, USA
mcarl6@kent.edu

Sheng Lu
TU Darmstadt, Germany
boblu@ruc.edu.cn

Ali Al-Ramadan
Kent State University, USA
aalramd@kent.edu

Abstract

We report an experiment in which we use machine learning to validate the empirical objectivity of a novel annotation taxonomy for behavioral translation data. The HOF taxonomy defines three translation states according to which a human translator can be in a state of Orientation (O), Hesitation (H) or in a Flow state (F). We aim at validating the taxonomy based on a manually annotated data-set that consists of six English-Spanish translation sessions (approx 900 words) and 1813 HOF-annotated Activity Units (AUs). Two annotators annotated the data and obtain high average inter-annotator accuracy 0.76 (kappa 0.88). We train two classifiers, a Multi-layer Perceptron (MLP) and a Random Forest (RF) on the annotated data and tested on held-out data. The classifiers perform well on the annotated data and thus confirm the epistemological objectivity of the annotation taxonomy. Interestingly, inter-classifier accuracy scores are higher than between the two human annotators.

1 Introduction

Translation is considered to involve complex and non-linear cognitive processes (Krings, 2001). Understanding the intricacies of the temporal dynamics of these processes is a fundamental aspect in Translation Process Research (TPR).

Various approaches have been proposed over the past 40 years to understand the distinct phases and mental states experienced by translators (Jakobsen, 2017). Starting with Think-Aloud Protocols in the 1980s, in which translators comment their own translation behavior during their translations (Königs, 1987; Krings, 2001), the field of enquiry has moved towards less invasive technologies, that is, keystroke logging and eye tracking (Hvelplund, 2016; Carl et al., 2016). The recordings of these logging tools make it possible to assess the flow of translation in a seamless way and to investigate how translations evolve in time; where translators type smoothly, where they get stuck, and where they search for (external) resources, etc.

One approach to analysing the translation process has been to segment the behavioral Translation Process Data (TPD) into processing units (Alves and Vale, 2009; Schaeffer et al., 2016). But how these segments should be defined and what they exactly represent has been a topic of continuous exploration and debate. The assessment of the translation rhythm (aka ”Pause Analysis” (Kumpulainen, 2015; Muñoz and Apfelthaler, 2022)) has provided valuable insights into translation patterns as produced by more or less experienced translators (Jakobsen, 2011), for different levels of text complexity (Hvelplund, 2016), for different translation goals (Zou et al., 2022b), post-editing behavior (Jia et al., 2019) and also for spoken translation (e.g., interpretation, sight translation, (Zou et al., 2022a)). The underlying assumption has been that longer keystroke pauses are indicative of more challenging translations, while stretches of smooth typing can be observed when there are no/less translation hurdles or difficulties (Lacruz et al., 2014). However, determining an exact pause threshold to differentiate these phenomena remains a challenge. Many studies (Krings, 2001; O’Brien, 2006; Kumpulainen, 2015; Vieira, 2016, among others) present varying segmentation methods with different pause thresholds, ranging from 300ms to five seconds or more. These strands of research rely on deterministic fragmentation techniques to segment the key logging and eye tracking data into Translation Units (TUs, (Alves and Vale, 2009; Carl and Schaeffer, 2017)) or Activity Units (AUs, (Hvelplund, 2016; Schaeffer et al., 2016)). However, these approaches lack intuitive labeling and thus make it difficult to derive a comprehensive understanding of the complex nature of the translation process and how it unfolds over time. Some researchers suggest a hierarchical process model (Schaeffer and Carl, 2013) and others (Muñoz and Apfelthaler, 2022; Dragsted, 2010) advocate a translator-specific fragmentation of the TPD into processing units depending on the translators' typing speed.

Combined, this suggests that human translation processes are embedded in a hierarchical mental architecture, encapsulating various processing strata. A hierarchical approach to understanding translation has the potential to offer a nuanced understanding of translators' behavior and strategies on various interacting levels of analysis. In order to advance this project, a novel higher-level segmentation taxonomy was introduced in (Carl et al., 2024) that fragments the TPD in three broad phenomenal translation states, Hesitation (`H`), Orientation (`O`) and Translation Flow (`F`). The `HOF` taxonomy assumes that behavioral traces of these three states can be observed in the TPD and that translators can be at any one point in time in only one of the three states.

In previous work (Carl et al., 2024) we have annotated a small corpus with `HOF` translation states. The corpus is publicly available as part of the CRITT TPR-DB[1]. The `HOF` annotation corpus provides a layer of manual annotation, introducing segment labels of an assumed phenomenal layer of translation processes, suited to analyse the hierarchical embedding of translation processes.

However, the `HOF` taxonomy, capturing qualities of conscious translator experience, is entirely new territory and the genaralizability and validity of the annotation schema is still unclear. In this paper, we therefore conduct further investigation to assess whether and to what extent this new taxonomy is epistemologically valid — that is, we want to investigate to what extent different annotators might agree the `HOF` states to represent a phenomenal translation "reality". While a manual annotation has shown a varying amount of agreement between two annotators (Kappa 0.37 and 0.88, (Carl et al., 2024), in this paper we use ML techniques to further validate the consistency of the taxonomy.

AI and ML techniques can be used in various ways and for different purposes. Mollo (2024), for instance, enumerates a few scenarios where AI can be used as: AI-as-engineering (industrial and commercial projects, as e.g., MT systems), AI-as-psychology to improve our understanding of biological intelligence, AI-as-idea or AI-as-recreation for recreating biological intelligence in artificial systems. AI can also be used for "exploring intelligence spaces" so as to uncover new forms of intelligence that are different from human intelligence or to uncover algorithms (Zhong et al., 2023).

In this paper we use ML techniques to investigate the "epistemological objectivity" (Searle, 1998; Searle, 2017, see also section 5 for a discussion) of the `HOF` annotation schema. That is, we are interested in verifying whether ML can reproduce the results of our manual annotations to a similar amount of accuracy. We assume that if the trained models performs well on the classification task, it confirms the objectivity of the annotation taxonomy used to create the training data. Conversely, poor model performance would indicate issues with the annotation taxonomy.

For instance, it might be the case that, even though two annotators agree in their annotation label, they might be biased by some intuition, cultural or otherwise un-observable features which may not be accessible to the ML technology. However, if ML reaches similar results of accuracy as the inter-annotator agreement indicates, we take the annotation taxonomy to implement reproducible and epistemological objective annotation criteria. That is, as ML lacks subjective, personal or cultural influences in the process annotation process (i.e., `HOF` state labeling), high accuracy on held-out testing data may be an indicator of stable results with minimum bias,

In section 2 we describe the data and the manual annotation process of the reference (training) data. Section 3 describes our implementation of two classifiers — Multi-Layer perception (MLP) and Random Forest (RF) — while section 4 reports our training and evaluation on a set of 1813

[1]The annotations can be downloaded from here `http://critt.as.kent.edu:3838/public/State_Annotation_Phases.zip`

HOF-annotated AUs. We report higher precision and recall between the classifier and annotator as compared to inter-annotator agreement between the two annotators. Section 5 concludes with a discussion on Searle's notion of "ontological subjectivity" and "epistemological objectivity" and their relation to the evaluation of our HOF states.

2 Activity Units, Translation Units and Translation States

The translation process can be conceptualized as a dynamic flow of mental processes marked by information input as gathered through eye movements and textual output in the form of keystrokes or mouse movements. As translators navigate the source text (ST) and produce the target text (TT), their behavior is influenced by numerous factors. To better understand these processes, different approaches have been pursued that fragment translation-behavioral data (keystrokes and gaze data) into processing units. Figure 1 shows a progression graph that depicts three ways of segmenting the approximately 28 seconds of the plotted translation session. TUs (Carl and Kay, 2011; Alves and Vale, 2009), indicated in the top in Figure 1, are characterized by a typing pause (a blank space in top line of the Figure) followed by a typing burst (or Production Unit, PU) indicated as grey boxes. AUs (Schaeffer et al., 2016; Hvelplund, 2016) are constructed based on the coordination of gaze activities and typing behavior and are marked at the bottom in Figure 1 in different colors. Three distinct HOF translation states are indicated with black dotted lines (Carl et al., 2024). Boundaries of translation states coincide with AU boundaries, so that sequences of AUs can be used to fragment the TPD into HOF translation states. This section explores these constructs and the three annotations, emphasizing their significance and interplay.

2.1 Translation Units

TUs segment the continuous stream of translation activities (keystrokes) into stretches of typing and pausing. They capture the translator's perception and actions, indicating the challenges they encounter during the translation process (Malmkjaer, 1998). In Figure 1, TUs appear as successive pauses and typing bursts of fluent production (or PUs). The pauses that occur between PUs are taken to be indicators of elevated translation effort, as it is assumed that during those breaks translators engage in reflection or (mental) search (Dragsted, 2010). Sequences of TUs have been used to compute pause-word-ratio (Lacruz et al., 2014) as indicators of cognitive effort. However, TUs often lack the granularity to explain precisely what occurs during the pauses. Moreover, they do not differentiate whether a translator is directing their focus towards the ST or the TT during these intervals (Schaeffer et al., 2016).

AU	AU activity	Color	Effort	Effect
T1	ST reading	Blue	+	-
T2	TT reading	Green	+	-
T4	translation production	Yellow	-	+
T5	ST reading with concurrent production	Red	-	+
T6	TT reading with concurrent production	Dark Green	-	+
T8	no observed behavior for more than one second	Black	+	-

Table 1: Types of AUs, color code in Figure 1 and levels of translational effect and cognitive effort.

2.2 Activity Units

AUs provide a more fine-grained view on the translation process, focusing on the coordination of the translator's eyes and hands. It addresses some of the inherent limitations of TUs. In our data, we categorize AUs into six types as presented in Table 1 (Carl et al., 2016). They provide more detailed insights into how translators engage in various aspects of the translation process. The classification of AUs is based on whether translators are actively involved in translation production, reading the ST or TT, or simultaneously reading and writing. As shown in Table 1, each type of AU can be associated with a degree of translational effects (typing activities) and cognitive effort (i.e., gazing). For instance, an AU of type T1 indicates ST reading which results in low levels of effects (no translation is typed) but higher amounts of cognitive effort (mental resources are allocated). In simpler terms, it means that the translator primarily focuses on understanding the ST, with minimal to no simultaneous translation work, as depicted in Figure 1.

2.3 HOF Translation States

HOF Translation States offer insights into qualities of the translator's experience. Carl et al (2024) distinguish between three translation states: A state

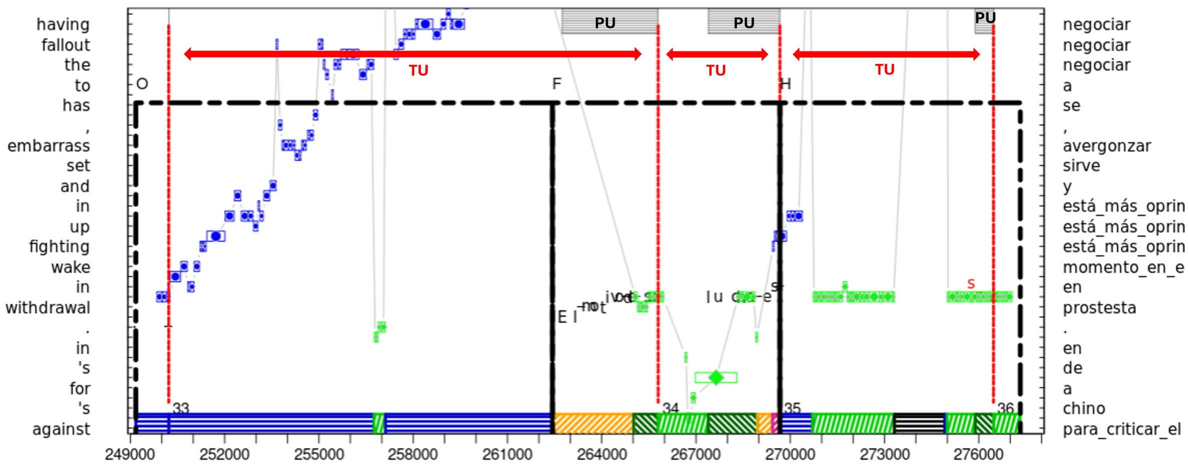

Figure 1: Progression graph of a small snippet of the translation session (BML12/P02_T3). Production time in milliseconds is indicated in the horizontal axis. Vertical axis refers to the ST on the left side and the TT on the right side. The blue dots and green diamonds represent eye movements on the ST and TT respectively. The black and red characters are insertion and deletion respectively. AUs are marked as colored bars on the bottom, TUs are indicated with red lines, and PUs as gray boxes in the top of the graph. Translation States are sequences of AUs, indicated by black dashed boxes, labeled `O`, `F`, `H`. The graph represents a segment of approximately 28 seconds (249.000ms - 277.000ms) of an English-to-Spanish translation.

of orientation (`O`) refers to the translator's behavior when feeling the need to get acquainted with the source text (ST). It is characterized by linear, forward-reading behavior of the ST. The Flow state (`F`) represents a phase in which the translator is immersed in translation production, generating the TT with ease and minimal interruption. It is marked by fluent translation production with minimal reading ahead and short pauses. A state of Hesitation (`H`) emerges out of surprise, where unexpected challenges prompt the translator to revise and re-read. This state indicates moments of uncertainty or cognitive challenge, signifying areas where the translator is challenged with complexities in the source text or struggles to find suitable translations. These distinct translation states are annotated in the progression graph in Figure 1, exemplifying associated typical behavioral correlates.

2.4 Empirical Data

We use a set of six translation sessions from the CRITT TPR-DB that were previously annotated with `HOF` translation state labels (Carl et al., 2024). The CRITT TPR-DB (Carl et al., 2016) is a collection of currently more than 5000 translation sessions, amounting to hundreds of hours of TPD, that is compiled into a consistent publicly available database. The CRITT TPR-DB is extensively documented in numerous publications and summary tables with more that 300 product and process features are available in a compiled form.[2]

In this study we use six English-to-Spanish translation sessions from BML12[3] The BML12 study consists of 184 translation sessions by Spanish translation students that were recorded in 2012 in Copenhagen and in Spain (Barcelona). The `HOF` annotation taxonomy was developed based (among others) on six BML12 sessions and annotated in 2022 by two advanced (Chinese and Japanese) translators. A special purpose interface was used to annotate the translation sessions, similar to Figure 1. The annotation process is described in detail (Carl et al., 2024). The six annotated sessions consist in total of 42 segments (sentences) with 854 source words. The translations of these 42 segments resulted in 1813 AUs which were annotated with `HOF` labels. In this study we used the 1813 `HOF`-annotated AUs for training and evaluating two classifiers.

[2]See the CRITT website `https://sites.google.com/site/centretranslationinnovation/tpr-db`The TPD can be downloaded free of charge from sourceforge `https://sourceforge.net/projects/tprdb/`, an introduction to the usage and a free trial account is provided here `https://sites.google.com/site/centretranslationinnovation/tpr-db/getting-started`

[3]The MultiLing data and BML12 study is described: `https://sites.google.com/site/centretranslationinnovation/tpr-db/public-studies#h.p_iVVuCQOHJx2O`

2.5 Manual Annotation

As reported in (Carl et al., 2024), the manual annotation involved five phases: Phases 1, 2, and 4 were trial annotations and are not considered here. In Phase 3, 1288 AU were annotated, but the absence of a structured approach resulted in a Kappa score of 0.37, indicating a moderate agreement between the two annotators (see Table 2). In Phase 5, a structured approach with a decision tree and guidelines was defined (Carl et al., 2024), resulting in a significant improvement in inter-annotator agreement, as shown in the high Kappa score of 0.88. Table 2 shows the inter-rater accuracy and Kappa scores along with the number of AUs used in Phase 3 and 5.

Phase	Total AUs	Kappa	Accuracy
3	1288	.37	.66
5	525	.88	.93

Table 2: Kappa scores for Phases 3 and 5 of annotation. Phase 3 involves five sessions, while Phase 5 involves session P04_T2 of the BML12 study. The average accuracy for all 1813 annotations is .74.

Furthermore, Table 3 offers a breakdown of the number of AUs for each of the three translation states in Phases 3 and 5, and for both annotators (Y and T). There is a noticeable shift in the distribution of annotated AUs across these states between the two phases. In Phase 3, the difference in the numbers of AUs annotated by T and Y across the three states suggests distinct interpretations of the states. This is the reason for the low Kappa score of 0.37 and low Accuracy of 0.66 in Table 2. The elaboration of a decision tree and annotation guidelines prior to Phase 5 clearly leads to a better alignment between the two annotators, evidenced not only by their closely matching counts across states but also by the high Kappa score of 0.88 (Accuracy 0.93). In our experiments, we use annotations from Phases 3 and 5 as a training/test corpus in section 4. Given the amount of coordination and mutual adjustment of the two annotators, we decided to corroborate the empirical objectivity of the annotation schema using MT.

3 Training Translation State Classifiers

In this section we describe two classifiers that were used to assess the annotated translation states. While the variation of inter-annotator agreement, as reported in Table 2, indicates that annotators

	Phase 3		Phase 5	
State	Y	T	Y	T
`H`	403	331	216	217
`O`	275	108	51	55
`F`	610	849	258	253
total	1288	1288	525	525

Table 3: Number of AUs in the two AU annotation phases for the two annotators T and Y.

are able to learn and agree on annotation guidelines and to generalize and reproduce the underlying concepts, this does not necessarily mean that those generalizations can also be learned and reproduced by a ML classifier. If, however, ML techniques can reproduce manual annotations with high accuracy we can be more certain about the "epistemological objectivity" (Searle, 2017) of the annotation schema. Besides, once a classifier is trained, we will also be able to automatically annotate new data. Therefore, in this study we only describe an evaluation of the trained classifier on the manually annotated data and leave a full-blown analysis on a large corpus for future research.

3.1 Multi-layer Perceptron

An MLP is a supervised simple feed-forward neural network. It consists of multiple layers, and each layer is fully connected to the following one. Figure 2 shows the structure of a two-layer MLP.

For our task, there are 34 cells in the input layer, and each of them corresponds to one of the 34 AU features (see Appendix A). There are 3 cells in the output layer corresponding to 3 state labels {`H`, `O`, `F`}. The output is a set of probabilities, each of which represents the probability that an input is classified as a certain state. The final prediction is the state with the highest probability.

We implemented MLP classifiers using `sklearn`[4]. We set the following parameters in `MLPClassifier`:

- `hidden_layer_sizes`: number of neurons in hidden layers
- `batch_size`: size of mini-batches
- `max_iter`: maximum number of iterations
- `learning_rate_init`: learning rate used
- `random_state`: random seed
- `solver`: weight optimizer

[4]See https://scikit-learn.org.

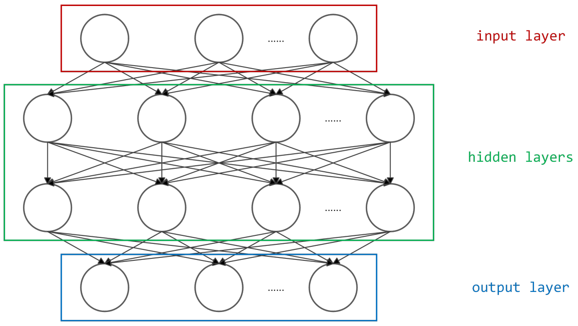

Figure 2: The structure of a two-layer MLP.

		Best		Average	
		acc.	F1	acc.	F1
T	MLP	85	69	67	52
	RF	85	75	78	64
Y	MLP	79	73	66	51
	RF	83	76	75	57

Table 4: Best and average accuracy (acc.) and F1-score (F1, in percentage) for the RF and MLP classifiers for both annotators.

3.2 Random Forest

We implemented an RF classifier using `sklearn` (see footnote 4). RF is a supervised ML method that trains a model on the annotated ('gold') data. An RF is a set of decision trees where the output of the classifier is the class selected by most trees (majority vote). In our trials we set this number to 500. RFs are said to be rather robust with respect to variations in the data, as the entry point to each of the decision trees varies probabilistically, while the algorithm averages over the differences. New, unseen data can, therefore, be classified reliably. Another advantage of RFs is that the importance of the features can be ranked, which may provide helpful insights into feature design. In Appendix B we provide the ranking of feature importance for the 34 AU features and the two models trained on the annotated Y and T data.

4 Evaluation of Classifiers

The training of AU-to-state classifiers is based on six annotated sessions from Phase 3 and 5 with 1813 datapoints (AUs), as shown in Table 3.

We trained classifiers based on two backbone models: MLP and an RF. The annotated data are split into 70% for training (1269 data points), and 30% for testing (544 data points). We used the 34 features that are described in the Appendix A in Table 9 for the classifiers.

In a 10-fold cross validation with RF and MLP we get best average accuracy values of 0.85 for the two classifiers.Table 4 shows the best and average performance of classifiers based on the two models. We observe the best accuracy and F1-score for the RF classifiers and annotator T.

4.1 Multi-layer Perceptron

For the MLP classifier, we used the Adam optimizer (Kingma and Ba, 2014) with an alpha value of 1e-5. The MLP architecture consisted of three hidden layers with 400, 200, and 400 units in size, respectively. We set the random state to 1 for reproducibility. The classifiers were trained separately on the data annotated by annotators Y and T. We also used the `StandardScaler` from `sklearn`, which standardizes features by removing the mean and scaling to unit variance.

4.2 Scaling Data

It is worth noting a difference in the performance of the MLP classifiers when trained on scaled data vs. non-scaled data. The effect of `StandardScaler` to the performance of the MLP classifier is significant. When the MLP model is trained on unscaled data, its average accuracy drops significantly. However, the scaler does not impact the results of the RF classifier.

MLP	State	Prec.	Rec.	F1	Support
T	H	0.77	0.79	0.78	164
	O	0.72	0.69	0.70	45
	F	0.91	0.90	0.91	335
Y	H	0.76	0.74	0.75	204
	O	0.66	0.59	0.63	96
	F	0.81	0.86	0.83	244

RF	State	Prec.	Rec.	F1	Support
T	H	0.82	0.83	0.82	168
	O	0.72	0.48	0.57	48
	F	0.90	0.94	0.92	328
Y	H	0.84	0.84	0.84	189
	O	0.84	0.73	0.78	93
	F	0.88	0.92	0.90	262

Table 5: Precision (Prec.), Recall (Rec.), and F1-score (F1) for MLP (top) and RF (bottom) classifiers trained on 1269 AU annotations and evaluated on a test set of 544 AU annotations for both annotators.

4.3 Precision and Recall

A fine-grained assessment of the classification report in Table 5 reveals that values for Precision, Recall and F1-score are differently distributed for the three States. State `F` has highest precision and recall values for both annotators Y and T and for both classifiers. As noted in (Carl et al., 2024), it is comparatively easy to detect this state in the behavioral data and it also has the best inter-rater agreement, as discussed in section 2 and Table 3. States `O` and `H` seem to be more difficult to separate and may be easily confused. Note that that Table 5 shows this to be the case for both annotators, T and Y.

4.4 Comparing Y and T labels

Table 6 shows two confusion matrices between the Y annotations and the RF predictions (on the left) and the T annotations (on the right). As the two matrices show, as well as indicated in Tables 3 and 5, the distribution of states are unequally distributed. There are almost three times more AUs with `F` label, as compared to `O` states.

A large number of states `H` seem to be classified as `F`, which indicates that more distinctive features might need to be developed, so as to better distinguish between states `F` and `H`.

True	RF Predictions			True T-labels		
Y-labels	H	O	F	H	O	F
H	156	13	20	77	38	74
O	14	67	12	26	37	30
F	15	2	245	16	7	239

Table 6: Left: Confusion Matrix for predictions of Y-labels of the test set for RF classifier shown in Table 5 (bottom). Right: Confusion Matrix for the same test set but against the true T-labels for the same data points.

The confusion matrices in Table 6 show that predictions produced by the RF classifier trained on the annotated Y-data correspond to a higher degree with the same annotator than the labels between the two annotators. This may indicate that each annotator is consistent in itself, whereas larger disagreement can be observed between the annotators. For instance, in the upper row, out of the 189 Y-annotated `H` labels, 156 labels were corrected predicted by the trained RF, 13 were predicted as `O` and 20 as `F` states. In contrast, annotators Y and T agree in `H` label only 78 cases. AUs that annotator Y considers `H` receive in 38 instances label `O` and 74 cases the label `F` by annotator T.

4.5 Accuracy across Annotators and Classifiers

In order to corroborate the assumptions in the previous subsection, we assess accuracy patterns across the two classifiers and annotators in more detail. We trained the RF (R) and MLP (M) classifiers with a training set of 1269 AUs to predict T and Y labels, as outlined in section 3. This provided us with four models for the two classifiers (M and R) and two annotators (T and Y). Successively, each of the four models (MT, MY, RT, and RY) predicted a list of state labels for the 544 examples in test set. We thus obtain six lists of state label predictions for the test set: four lists of predictions from the four classifiers (MT, MY, RT, and RY) and in addition the original labels from the manual T and Y annotations. Table 7 shows accuracy scores for the 6 $\times$ 6 pair-wise combinations of these label lists. Since accuracy scores are symmetrical (i.e., Accuracy(x, y) == Accuracy(y, x)), Table 7 only shows the lower part of the rectangular matrix. Note also that Accuracy(x, x) == 1 and that the triangle below the diagonal adds up to 15 accuracy pairs (cells).

	T	Y	RT	RY	MT
Y	.76	1	—	—	—
RT	.87	*.72*	1	—	—
RY	*.77*	.86	.78	1	—
MT	.85	*.71*	**.91**	.77	1
MY	*.73*	.80	.74	**.85**	.76

Table 7: Accuracy scores for different pairs of Classifiers (R and M) and Annotators (T and Y.

The accuracy scores in Table 7 range between .71 and .91. As discussed in sections 4.1, higher accuracy scores are observed for RF than for MLP and for annotator T as compared to annotator Y. However, contrary to what one might expect, the highest accuracy scores are obtained between predictions of two classifiers trained on the same data, rather than between predictions of a classifier and the data population it was trained on. Thus, Table 7 reveals that:

1. highest accuracy scores are observed when comparing predictions of two different classifiers trained on data of the same annotator. Thus the two comparisons: MT/RT and MY/RY produce among the highest accuracy scores of .91, and .85 respectively. These numbers are marked in **bold** in Table 7

2. high accuracy scores, but not quite as high, are also observed when comparing predictions of a classifier evaluated against the manual annotations of the same annotator that the classifier was trained on. Thus the the four accuracy scores: RT/T, RY/Y, MT/T, and MY/Y produce the second highest accuracy scores of .87, .86, .85, and .80 respectively. This is the case discussed in the context of Table 6 (left).

3. as can be expected, the predictions of two classifiers trained on different annotators provides lower accuracy values as compared to those in item 1. and 2 above. These pairs of `HOF` state label predictions have the following accuracy values: MY/RT:.74, RY/RT:.78, MT/RY:.77, and MY/MT:.76.

4. surprisingly, even lower is the accuracy between the two manual annotations T/Y. With a value of .76 it is just slightly higher than the accuracy values of a manual annotation and a different classifier in item 5. This is the case discussed in the context of Table 6 (right).

5. the lowest accuracy scores are observed when comparing predictions of a classifier that was trained on one annotator A but evaluated with a manual annotation of the other annotator B. The the four comparisons: MT/Y, RT/Y, MY/T, and RY/T produce the lowest accuracy scores of .71, .72, .73, .77 respectively. These numbers are marked in *italics* in Table 7

The results are somewhat puzzling. Most surprising is perhaps the finding that the output of the classifiers in item 1. are more consistent (higher accuracy) than the the classifiers in 2 and that accuracy values in 3. are higher than inter-rater accuracy in 4.

Provided that a(ny) classifier generalizes and approximates the inherent structure of the manual annotations, there will be some noise in the generalizations. Under this assumption we expect that accuracy values in 2. should be higher than in 1, since the noise of two classifiers (in item 1) would multiply. Presumably, each of the two classifiers (M and R) would 'infer' their own generalizations which, we would assume, are likely less compatible than the classifier's own generalization about the set of manual annotations which the classifier was trained on (as in item .2). Provided that the manual annotations are consistent, i.e., they are 'gold' data, why then do pairs of classifiers trained on the same data produce higher accuracy values as compared to the gold data?

Why would it be the case that predictions from two different models (R and M) produce more consistent predictions as each of the classifiers evaluated against the test data taken from a population that they were trained on? Provided the test data is correct, how is it possible that, despite their very different nature and implementation, RF and MLP make similar but wrong predictions?

Similar surprising is the observation that T/Y inter-annotator accuracy in item 4. is lower than the predictions of the classifier trained and evaluated on two different annotators in item 3.

Also this outcome suggests that the two classifiers may have inferred similar generalizations that somehow capture similarities between the T and Y training sets, but that do not, however, account correctly for the structure of the test set. This idea is corroborated in the accuracy values reported in item 5. which shows that the worst values are obtained by evaluating a classifier on a manual test set of a different annotator.

The results indicate that an evaluation of the classifier on manually annotated data or on automatically generated test sets may lead to different results. The results may also indicate that the training set is perhaps not sufficiently large to capture the instances that are represented in the test set. However, we take it that our experiments validate the epistemological objectivity of the `HOF` taxonomy, as the classifier perform well on the task at hand.

5 Discussion and Conclusion

Human translation is a complex cognitive process that involves numerous interacting processes. To understand and analyse these processes, one approach to Translation Process Research (TPR) has been to collect and synchronize behavioral data (keystrokes and gaze data) from translation sessions and to segment the flow of data into various kinds of processing units. Several automatic segmentation approaches have been suggested, but as the labels often lack intuitive understanding it is difficult to interpret the data.

A novel higher-order `HOF` taxonomy has been proposed (Carl et al., 2024) that segments the data into three phenomenal states in which a translator can be: a state of orientation (`O`) accounts for the

	Ontology	Epistemology
Subjective	EXISTENCE OF THE SUBJECTIVE • Reality as I experienced it (intentions, attitudes, pain, beliefs, desires, etc) • Conscious personal experience	KNOWLEDGE OF THE SUBJECTIVE • Reality as it is judged by me (opinions, preferences, etc) • What "I" know to be the case
Objective	EXISTENCE OF THE OBJECTIVE • Reality as it exists: physical, spatial, temporal (mountains, molecules, etc.) • Exists independent of perception	KNOWLEDGE OF THE OBJECTIVE • Reality as ”we” describe it: norms, conventions (money, marriage, etc.) • Assertions "we" make about reality

Table 8: Modes of existence according to Searle.

need of ST information input which is characterized by reading-ahead in the ST. In a flow state (F), translations are fluently produced, and the state of hesitation (H) reflects surprise or uncertainty, which is characterized by regressions, re-fixations and text modifications. Together with the HOF taxonomy, (Carl et al., 2024) specify a decision tree that provided criteria for the annotation process.

A small corpus of behavioral data annotated and released (Carl et al., 2024). The annotated data consists of six English-Spanish translation sessions (approximately 900 words) and 1813 HOF-state annotated Activity Units (AUs, (Carl et al., 2016)). Two annotators annotated the data with HOF labels and — after specifying a decision tree and annotation guidelines — annotators reached a good inter-rater agreement.

Given the novelty of the annotation taxonomy, we investigate how well the HOF annotations can be reproduced. We use machine learning (ML) classifiers to validate the ”epistemological objectivity” (Searle, 2017) of the annotation schema. That is, we deploy a Multi-layer Perceptron and a Random Forests classifier to assess the ”objectivity” of the manual annotations, where high accuracy of the ML classifiers would indicate the validity of the underlying HOF annotations taxonomy.

In his discussion about ”modes of existence”, (Searle, 1998; Searle, 2017) makes a distinction between, on the one hand, subjective and objective ways of understanding and, on the other hand, between the epistemology and the ontology of knowledge and reality (see Table 8). Whereas ontology is a branch of metaphysics that deals with the nature of being, epistemology is the branch of philosophy concerned with the theory of knowledge.

Ontological subjectivity then refers to the idea that subjective experiences is a form of reality, but there may not be an independent, objective reality beyond these subjective constructions (see Table 8). Epistemological objectivity is the idea that certain aspects of reality can be known objectively, independent of my beliefs, perspectives, or interpretations. Objective knowledge can be discovered or verified through rational inquiry, observation, or evidence, regardless of subjective opinions or interpretations.

Despite the fact that consciousness has a subjective mode of existence—and is thus not directly accessible to scientific inquiry—Searle claims that this does not prevent us from having an epistemological objective science of consciousness. While translators experience subjective states of orientation, hesitation and flow, these states, we assume, can be recovered in the behavioral data and studied under epistemically objective conditions. Norms, regulations or—as in our case the HOF annotation taxonomy—can be understood, deployed and objectively verified within observable TPD in a given context. Our results suggest, however, that there might be a gradual slope between epistemological subjective and epistemological objective modes of existence, rather than a binary one. Table 7 suggests that, despite a well-formulated HOF state decision tree as described in (Carl et al., 2024), a perfect agreement between different annotators may not always be possible[5]. Accuracy values,

[5]Similar findings have been reported in countless translation

such as those in Table 7, may thus provide an index for the degree of epistemological objectivity, where higher accuracy values indicate greater epistemological value of the underlying taxonomy (or norm), and thus allow for higher objectivity while lower accuracy values indicate increased possibilities for epistemological subjectivity. Surprisingly, then, our findings indicate that the two different classifiers (MT/RT and MY/RY) trained on the same data are able to arrive at higher epistemological objectivity as compared to the two human annotators who follow the same annotation guidelines. It suggests that different classifiers are able to generalize the (training) data in a similar way which, however, deviates from generalizations that our annotators from the annotation guidelines and decision trees.

Acknowledgements

The authors would like to thank Masaru Yamada and Yuxiang Wei for extended discussions about some of the topics addressed in this paper.

References

[Alves and Vale2009] Alves, Fabio and D. Vale. 2009. Probing the unit of translation in time: Aspects of the design and development of a web application for storing, annotating, and querying translation process data. *Across Languages and Cultures 10(2)*, pages 251–273.

[Carl and Kay2011] Carl, Michael and Martin Kay. 2011. Gazing and typing activities during translation : A comparative study of translation units of professional and student translators. *Meta 56(4))*, pages 952–975.

[Carl and Schaeffer2017] Carl, Michael and Moritz Schaeffer. 2017. Sketch of a noisy channel model for the translation process. In *Empirical Modelling of Translation and Interpreting*, pages 71–116. Berlin: Language Science Press.

[Carl et al.2016] Carl, Michael, Srinivas Bangalore, and Moritz Schaeffer. 2016. New directions in empirical translation process research. *Heidelberg: Springer International Publishing Switzerland. doi*, 10:978–3.

[Carl et al.2024] Carl, Michael, Yuxiang Wei, Sheng Lu, Longhui Zou, Takanori Mizowaki, and Masaru Yamada. 2024. Hesitation, orientation, and flow: A taxonomy for deep temporal translation architectures. *Ampersand*, page 100164.

evaluation studies.

[Dragsted2010] Dragsted, Barbara. 2010. Coordination of reading and writing processes in translation. In Shreve, Gregory M. and Erik Angelone, editors, *American Translators Association Scholarly Monograph Series*, volume 381, pages 41–62. John Benjamins Publishing Company.

[Hvelplund2016] Hvelplund, Kristian. 2016. Cognitive efficiency in translation. In Martín, Ricardo Muñoz, editor, *Reembedding Translation Process Research*, pages 149–170. John Benjamins Publishing Company, 9.

[Jakobsen2011] Jakobsen, Arnt. 2011. Tracking translators' keystrokes and eye movements with translog. In Alvstad, C., A. Hild, and E. Tiselius, editors, *Methods and Strategies of Process Research*, pages 37–55. John Benjamins Publishing Company.

[Jakobsen2017] Jakobsen, Arnt Lykke. 2017. Translation process research. In Schwieter, John W. and Aline Ferreira, editors, *The Handbook of Translation and Cognition*. Willey.

[Jia et al.2019] Jia, Yanfang, Michael Carl, and Xiangling Wang. 2019. How does the post-editing of neural machine translation compare with from-scratch translation? a product and process study. *The Journal of Specialised Translation*, 31:60–86.

[Kingma and Ba2014] Kingma, Diederik P and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Krings2001] Krings, H.P. 2001. *Repairing Texts. Empirical Investigations of Machine Translation Post-editing Processes*. Kent, Ohio: Kent State U.P.

[Kumpulainen2015] Kumpulainen, Minna. 2015. On the operationalisation of 'pauses' in translation process research. *The International Journal for Translation & Interpreting Research*, 7:47–58.

[Königs1987] Königs, Frank. 1987. Was beim Übersetzen passiert: theoretische aspekte, empirische befunde und praktische konsequenzen. *Die neueren Sprachen 2*, 2(86):162–185.

[Lacruz et al.2014] Lacruz, Isabel, Gregory Shreve, and M. 2014. Pauses and cognitive effort in post-editing. In O'Brien, Sharon, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia, editors, *Post-editing of Machine Translation: Processes and Applications*. Cambridge Scholars Publishing.

[Malmkjaer1998] Malmkjaer, K. 1998. Unit of translation. *Routledge Encyclopedia of Translation Studies*, page 286–288.

[Mollo2024] Mollo, Dimitri Coelho. 2024. Ai-as-exploration: Navigating intelligence space.

[Muñoz and Apfelthaler2022] Muñoz, Ricardo and Matthias Apfelthaler. 2022. A Task Segment Framework to study keylogged translation processes. *Translation and Interpreting*, 14.

[O'Brien2006] O'Brien, Sharon. 2006. Pauses as indicators of cognitive effort in post-editing machine translation output:. *Across Languages and Cultures*, 7(1):1 – 21.

[Schaeffer and Carl2013] Schaeffer, Moritz and Michael Carl. 2013. Shared representations and the translation process. a recursive model. *Translation and Interpreting Studies 8 (2)*, pages 169–190.

[Schaeffer et al.2016] Schaeffer, Moritz, Michael Carl, Isabel Lacruz, and Akiko Aizawa. 2016. Measuring cognitive translation effort with activity units. In *Proceedings of EAMT*. EAMT.

[Searle1998] Searle, John. 1998. *Mind, Language, and Society*. Basic Books.

[Searle2017] Searle, John. 2017. Reflections on free will, language, and political power, talks at google.

[Vieira2016] Vieira, Nunes. 2016. *Cognitive Effort in Post-Editing of Machine Translation: Evidence from Eye Movements, Subjective Ratings, and Think-Aloud Protocols*. Ph.D. thesis, Newcastle University.

[Zhong et al.2023] Zhong, Ziqian, Ziming Liu, Max Tegmark, and Jacob Andreas. 2023. The clock and the pizza: Two stories in mechanistic explanation of neural networks.

[Zou et al.2022a] Zou, Longhui, Michael Carl, and Jia Feng. 2022a. Patterns of attention and quality in english-chinese simultaneous interpreting with text. *International Journal of Chinese and English Translation & Interpreting*.

[Zou et al.2022b] Zou, Longhui, Michael Carl, Masaru Yamada, and Takanori Mizowaki. 2022b. Workshop on empirical translation process research. In *Proficiency and External Aides: Impact of Translation Brief and Search Conditions on Post-editing Quality*, page np. AMTA.

Appendix

A Features of Classifiers

Both classifiers were trained with a list of 34 features, shown in Table 9. The first 15 features, above the double line, prefixed with "TU_", are copied from the TU of which the AU is part (see Figure 1). These features thus encode the context of the AU. All "TU_" features relate to behavioral data, concerning the gaze and the keystroke data, and their duration.

The lower 19 features were extracted from and describe properties of AUs. Similarly, most of the AU features characterize the behavioral data within one AU. However, four of these features are related to properties of the translation product and four features include contextual from surrounding

Feature	Description of feature
TU_logDurTU	log-transformed duration of the TU
TU_WinSwitch	Number of gaze switches between ST and TT
TU_TrtT	Total reading time on the ST
TU_TrtS	Total reading time on the TT
TU_TrtST	ratio $\log((TrtS+1)/(TrtT+1))$
TU_TGset	Intersection of words IDs produced in next TU
TU_PauseDur	Ratio of $(Pause{+}1)/(DurTU{+}1)$
TU_ParTrtT	Duration of concurrent TT reading while typing
TU_ParTrtS	Duration of concurrent ST reading while typing
TU_ParFixT	#fixations during concurrent TT reading and typing
TU_ParFixS	#fixations during concurrent ST reading and typing
TU_InsDelLog	ratio of deletions and insertions $\log(Del+1)/(Ins+1))$
TU_FixT	Number of fixations on TT
TU_FixS	Number of fixations on ST
TU_FixDist	log of max. distance in Y-position of fixations on ST window (in pixel) $\log(FixSspanY+1)$
Type	Type of TU as described in Table 1
Gram5	concatenation of AU type with the preceding four AU types
Dur	Duration of the AU
SGnbr	#ST words for which translations were produced (concerns AU types T4, T5,T6)
TGnbr	#TT words produced (concerns AU types T4, T5,T6)
Ins	#Insertions (concerns AU types T4, T5,T6)
CrossS	Average Cross values for ST words produced in AU
CrossT	Average Cross values for TT words produced in AU
ProbSgaze	Average log probability of source words in GazePath
ProbTgaze	Average log probability of target Words in GazePath
ProbCgaze	Average log CrossS value in GazePath
ProbSTCgaze	Average log of joint ST, TT and CrossS value in GazePath
HSgaze	Average entropy of ST words in GazePath
HTgaze	Average entropy of TT words in GazePath
HCgaze	Average entropy of Cross values in GazePath
HSTCgaze	Average entropy of joint ST, TT and Cross in GazePath
Effort	sum of log duration for context AUs: T4, T5, T6
Effect	sum of log duration for context AUs: T1, T2, T8
Significance	*Effect* minus *Effort*

Table 9: List of features used for Classifier.

AUs information. The features *SGnbr* and *TGnbr* indicate how many source and target words were covered in the AU, while *CrossS* and *CrossT* are measures of the distance / reordering between the source and the target words (Carl et al., 2016). Four of the AU features refer to the nearby context of the AU. *Type* is type of AU (see Table 1) *Gram5* is the concatenation of AU type labels, while *Effort*, *Effect*, and *Significance* take into account Effort/Effect properties of the two surrounding AUs as described in Table 1.

In this study, we define *Effort*, *Effect*, and *Significance* for an AU to depend on the type and the duration of the two preceding two AUs and the next AU. The *Effort* of an AU is computed as the sum of *log(Dur(AU))* for each context-AU of Type T1, T2 or T8 (no keystroke activity is observed). The *Effect* is computed as the sum of *log(Dur(AU))* for each context-AU of Type T4, T5 or T6. The *Significance* of an AU is then its *Effect* minus its *Effort*, so that more *significant* AUs are characterized by longer stretches of text production.

B Importance of features in RF Classifier

Table 10 shows the 34 features in their order of importance as obtained during RF training. The list of features is ordered with respect to the importance of the T column (annotator T). The "N" column indexes features according to their importance for T, while the column header "O" provided the rank re-ordering of the importance for the Y data. There is a strong correlation between the two importance vectors of T and Y (R=0.95), indicating that slight differences in the annotation of T and Y do not seem to have a large impact on feature importance of the RT classification.

The context of AUs seems to be important for classifying their HOF label. Thus, the 15 TU-inherited features (those preceded by "TU_") make around 50% (49.18% and 49.94%) of the total importance for T and the Y respectively. Adding to this the importance of the features that account for the external context of the AUs, *Gram5*, *Effort*, *Effect* and *Significance*, increases the importance of context-related features to 72.16% and 73.87% respectively. That is, only 28% and 26% of the HOF state classification is due to AU internal characteristics. Those AU-local features are indicated in bold in Table 10. Also note that the first 11 most important features are all 'context' features which make up around 58% in the T set (57% in Y).

N	Feature	**T**	**Y**	O
1	*TU_PauseDur*	0.0829	0.0848	1
2	*Significance*	0.0770	0.0707	3
3	*Effect*	0.0709	0.0778	2
4	*Effort*	0.0569	0.0535	4
5	*TU_InsDelLog*	0.0513	0.0425	7
6	*TU_logDurTU*	0.0502	0.0345	10
7	*TU_FixS*	0.0431	0.0398	9
8	*TU_TrtS*	0.0421	0.0516	5
9	*Gram5*	0.0345	0.0278	13
10	*TU_FixDist*	0.0344	0.0271	14
11	*TU_TrtST*	0.0332	0.0467	6
12	***Dur***	0.0272	0.0293	11
13	*TU_FixT*	0.0271	0.0219	18
14	***Ins***	0.0268	0.0401	8
15	*TU_TrtT*	0.0246	0.0226	17
16	*TU_ParTrtS*	0.0229	0.0279	12
17	***CrossS***	0.0196	0.0250	15
18	*TU_WinSwitch*	0.0189	0.0154	24
19	***ProbCgaze***	0.0183	0.0146	27
20	*TU_ParTrtT*	0.0182	0.0234	16
21	*TU_TGset*	0.0181	0.0156	23
22	***Type***	0.0181	0.0152	25
23	***ProbSgaze***	0.0172	0.0134	34
24	*TU_ParFixT*	0.0166	0.0176	22
25	***HTgaze***	0.0162	0.0145	29
26	***HCgaze***	0.0161	0.0151	26
27	***HSTCgaze***	0.0160	0.0139	32
28	*TU_ParFixS*	0.0158	0.0204	20
29	***HSgaze***	0.0154	0.0143	30
30	***ProbTgaze***	0.0153	0.0136	33
32	***TGnbr***	0.0151	0.0208	19
31	***SGnbr***	0.0151	0.0203	21
33	***ProbSTCgaze***	0.0141	0.0140	31
34	***CrossT***	0.0110	0.0145	28

Table 10: Importance of features for T and Y annotations. Columns T and Y give the percentage for the respective features. Column N indicates the order of importance for the T annotator while O provides the order for Y annotator.